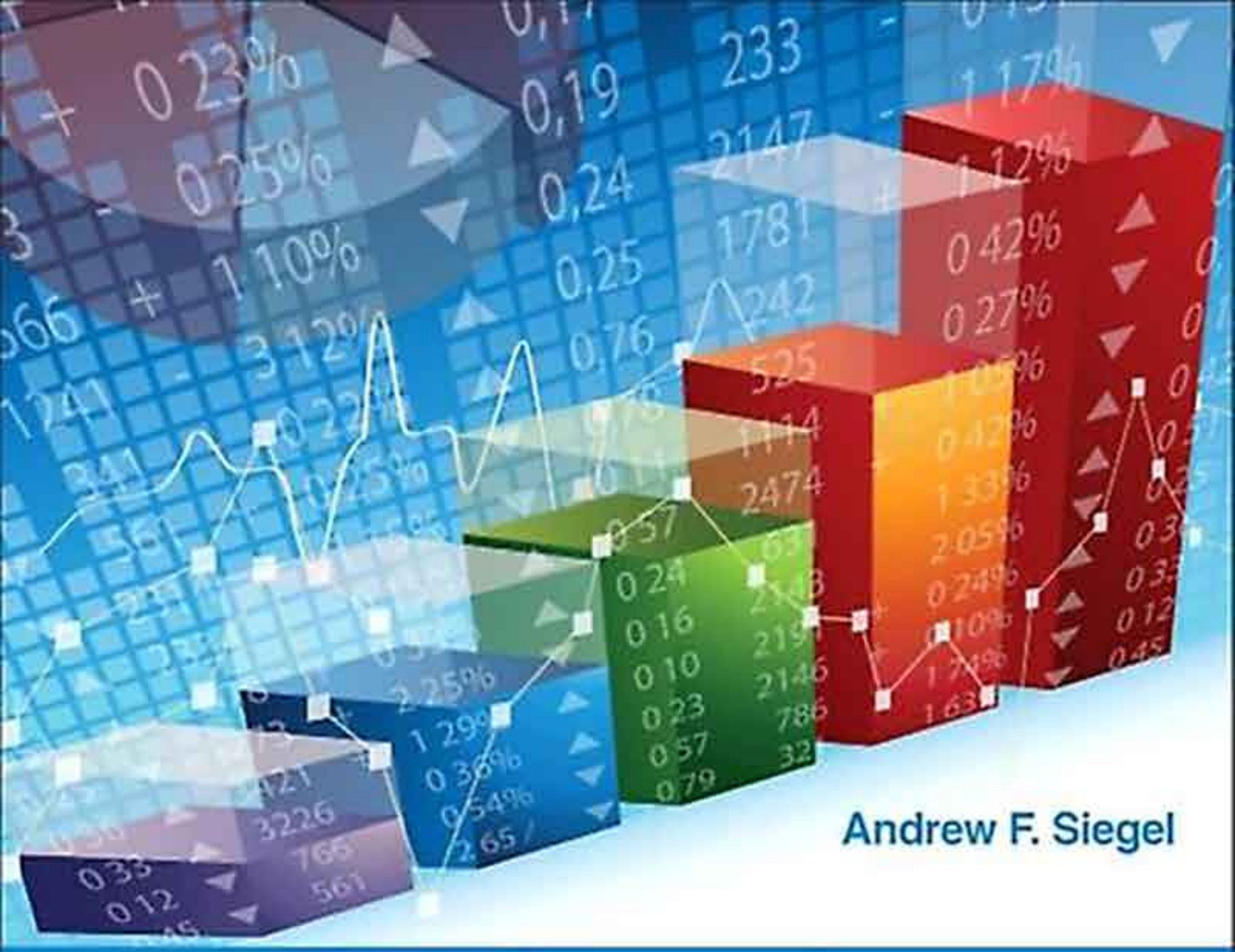

Andrew F. Siegel

Practical Business Statistics

Seventh Edition

Practical Business Statistics

Practical Business Statistics
Seventh Edition

Andrew F. Siegel
Department of Information Systems and Operations Management
Department of Finance and Business Economics
Department of Statistics
Foster School of Business
University of Washington
Seattle, USA

ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO
Academic Press is an imprint of Elsevier

Academic Press is an imprint of Elsevier
125 London Wall, London EC2Y 5AS, United Kingdom
525 B Street, Suite 1800, San Diego, CA 92101–4495, United States
50 Hampshire Street, 5th Floor, Cambridge, MA 02139, United States
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, United Kingdom

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

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

ISBN: 978-0-12-804250-2

Publisher: Nikki Levy
Acquisition Editor: Graham Nisbet
Developmental Editor: Susan Ikeda
Production Project Manager: Paul Prasad Chandramohan
Cover Designer: Mark Rogers

Typeset by SPi Global, India

To Ann, Bonnie, Clara, Michael, and Mildred

Preface

Statistical literacy has become a necessity for anyone in business, simply because your competition has already learned how to interpret numbers and how to measure many of the risks involved in this uncertain world. Can you afford to ignore the tons of data available (to anyone) online when you are searching for a competitive, strategic advantage? Humans are not born with an intuitive ability to assess randomness or process massive data sets, but fortunately there are fundamental basic principles that let us compute, for example, the risk of a future payoff, the way in which the chances for success change as we continually receive new information, and the best information summaries from a data warehouse. This book will guide you through foundational activities, including how to collect data so that the results are useful, how to explore data to efficiently visualize its basic features, how to use mathematical models to help separate meaningful characteristics from noise, how to determine the *quality* of your summaries so that you are in a position to make judgments, and how to know when it would be better to ignore the set of data because it is indistinguishable from random noise.

EXAMPLES

Examples bring statistics to life, making each topic relevant and useful. There are many real-world examples used throughout *Practical Business Statistics*, chosen from a wide variety of business sources, and many of them of recent interest (take a look at the status of Facebook, YouTube, and Google, relative to other top websites in Chapter 11, Figure 11.1.5). The donations database, which has the characteristics of 20,000 individuals, together with the amount that they contributed in response to a mailing, is introduced in Chapter 1 and used in many chapters to illustrate how statistical methods can be used for data mining with Big Data. The stock market is used in Chapter 5 to illustrate volatility, risk, and diversification as measured by the standard deviation, while the *systematic* component of market risk is summarized by the regression coefficient (a stock's "beta") in Chapter 11. Because we are all curious about the salaries of others, I have used top executive compensation in several examples and, yes, Enron was an outlier even before the company filed for bankruptcy and the CEO resigned. Quality control is used throughout the book to illustrate individual topics and is also covered in its own chapter (Chapter 18). Opinion surveys and election polls are used throughout the book (and especially in Chapter 9) because they represent a very pure kind of real-life statistical inference that we are all familiar with and use frequently in business. Using the Internet to locate data is featured in Chapter 2. Prices of magazine advertisements are used in Chapter 12 to show how multiple regression can uncover relationships in complex data sets, and we learn the value of a larger audience with a higher income simply by crunching the numbers. Microsoft's revenues and United States unemployment rates are used in Chapter 14 to demonstrate what goes on behind the scenes in time-series forecasting. Students learn better through the use of motivating examples and applications. All numerical examples are included in the Excel files on the companion website, with ranges named appropriately for easy analysis.

STATISTICAL GRAPHICS

To help show what is going on in the data sets, *Practical Business Statistics* includes over 300 figures to illustrate important features and relationships. The graphs are exact because they were drawn with the help of a computer. For example, the bell-shaped normal curves here are accurate, unlike those in many books, which are distorted because they appear to be an artist's enhancement of a casual, hand-drawn sketch. There is no substitute for accuracy!

EXTENSIVE DEVELOPMENT: REVIEWS AND CLASS TESTING

This book began as a collection of readings I handed out to my students as a supplement to the assigned textbook. All of the available books seemed to make statistics seem unnecessarily difficult, and I wanted to develop and present straightforward ways to think about the subject. I also wanted to add more of a real-world business flavor to the topic. All of the helpful feedback I have received from students over the years has been acted upon and has improved the book. *Practical Business Statistics* has been through several stages of reviewing and classroom testing. Now that

six editions have been used in colleges and universities across the country and around the world, preparing the seventh edition has given me the chance to fine-tune the book, based on the additional reviews and all the helpful, encouraging comments that I have received.

WRITING STYLE

I enjoy writing. I have presented the "inside scoop" wherever possible, explaining how we statisticians *really* think about a topic, what it implies, and how it is useful. This approach helps bring some sorely needed life to a subject that unfortunately suffers from dreadful public relations. Of course, the traditional explanations are also given here so that you can see it both ways: here is what we say, and here is what it means, all the while maintaining technical rigor.

It thrilled me to hear even some of my more quantitative-phobic students tell me that the text is actually *enjoyable to read!* And this was *after* the final grades were in!

CASES

To show how statistical thinking can be useful as an integrated part of a larger business activity, cases are included at the end of each of Chapters 3–12. These cases provide extended and open-ended situations as an opportunity for thought and discussion, often with no single correct answer.

ORGANIZATION

The reader should always know *why* the current material is important. For this reason, each part begins with a brief look at the subject of that part and the chapters to come. Each chapter begins with an overview of its topic, showing why the subject is important to business, before proceeding to the details and examples.

Key words, the most important terms and phrases, are presented in bold in the sentence of the text where they are defined. They are collected in the Keywords list at the end of each chapter and also included in the glossary at the back of the book (Hint! This could be very useful!). This makes it easy to study by focusing attention on the main ideas.

Extensive end-of-chapter materials are included, beginning with a *summary* of the important material covered. Next is the list of *key words*. The *questions* provide a review of the main topics, indicating why they are important. The *problems* give the student a chance to apply statistics to new situations. The *database exercises* (included in most chapters) give further practice problems based on the employee database in Appendix A. The *projects* bring statistics closer to the students' needs and interests by allowing them to help define the problem and choose the data set from their work experience or interests from sources including the Internet, current publications, or their company. Finally, the *cases* (one each for Chapters 3–12) provide extended and open-ended situations as an opportunity for thought and discussion, often with no single correct answer.

Several special topics are covered in addition to the foundations of statistics and their applications to business. Data mining with Big Data is introduced in Chapter 1 and is carried throughout the book. Because communication is so important in the business world, Chapter 13 shows how to gather and present statistical material in a report. Chapter 14 includes an intuitive discussion of the Box-Jenkins forecasting approach to time series using autoregressive integrated moving average (ARIMA) models. Chapter 18 shows how statistical methods can help you achieve and improve quality; discussion of quality control techniques is also interspersed throughout the text.

Practical Business Statistics is organized into five parts, plus appendices, as follows:

- Part I, Chapters 1 through 5, is titled "Introduction and Descriptive Statistics." Chapter 1 motivates by showing how the use of statistics provides a competitive edge in business and then outlines the basic activities of statistics and offers varied examples including data mining with Big Data. Chapter 2 surveys the various types of data sets (quantitative, qualitative, ordinal, nominal, bivariate, time series, etc.), the distinction between primary and secondary data, and use of the Internet. Chapter 3 shows how the histogram lets you see what is in the data set, which would otherwise be difficult to determine just from staring at a list of numbers. Chapter 4 covers the basic landmark summaries, including the average, median, mode, and percentiles, which are displayed in the box plot and the cumulative distribution function. Chapter 5 discusses variability, which often translates into *risk* in business terms, featuring the standard deviation as well as the range and coefficient of variation.
- Part II, including Chapters 6 and 7, is titled "Probability." Chapter 6 covers probabilities of events and their combinations, using probability trees both as a way of visualizing the situation and as an efficient method for computing probabilities. Conditional probabilities are interpreted as a way of making the best use of the information you have. Chapter 7 covers random variables (numerical outcomes), which often represent those numbers that are important to your business but are not yet available. Details are provided concerning general discrete distributions, the binomial distribution, the normal distribution, the Poisson distribution, and the exponential distribution.

- Part III, Chapters 8 through 10, is titled "Statistical Inference." These chapters pull together the descriptive summaries of Part I and the formal probability assessments of Part II, allowing you to reach probability conclusions about an unknown population based on a sample. Chapter 8 covers random sampling, which forms the basis for the exact probability statements of statistical inference and introduces the central limit theorem and the all-important notion of the standard error of a statistic. Chapter 9 shows how confidence intervals lead to an exact probability statement about an unknown quantity based on statistical data. Both two-sided and one-sided confidence intervals for a population mean are covered, in addition to prediction intervals for a new observation. Chapter 10 covers hypothesis testing, often from the point of view of distinguishing the presence of a real pattern from mere random coincidence. By building on the intuitive process of constructing confidence intervals from Chapter 9, hypothesis testing can be performed in a relatively painless, intuitive manner while ensuring strict statistical correctness (I learned about this in graduate school and was surprised to learn that it was not yet routinely taught in introductory courses—why throw away the intuitive confidence interval just as we are starting to test hypotheses?).
- Part IV, Chapters 11 through 14, is titled "Regression and Time Series." These chapters apply the concepts and methods of the previous parts to more complex and more realistic situations. Chapter 11 shows how relationships can be studied and predictions can be made using correlation and regression methods on bivariate data. Chapter 12 extends these ideas to multiple regression, perhaps the most important method in statistics, with careful attention to interpretation, diagnostics, and the idea of "controlling for" or "adjusting for" some factors while measuring the effects of other factors. Chapter 13 provides a guide to report writing (with a sample report) to help the student communicate the results of a multiple regression analysis to business people. Chapter 14 introduces two of the most important methods that are needed for time-series analysis. The trend-seasonal approach is used to give an intuitive feeling for the basic features of a time series, while Box-Jenkins models are covered to show how these complex and powerful methods can handle more difficult situations.
- Part V, Chapters 15 through 18, is titled "Methods and Applications," a grab bag of optional, special topics that extend the basic material covered so far. Chapter 15 shows how the analysis of variance allows you to use hypothesis testing in more complex situations, especially involving categories along with numeric data. Chapter 16 covers nonparametric methods, which can be used when the basic assumptions for statistical inference are not satisfied, that is, for cases where the distributions might not be normal or the data set might be merely ordinal. Chapter 17 shows how chi-squared analysis can be used to test relationships among the categories of nominal data. Finally, Chapter 18 shows how quality control relies heavily on statistical methods such as Pareto diagrams and control charts.
- Appendix A is the "Employee Database," consisting of information on salary, experience, age, gender, and training level for a number of administrative employees. This data set is used in the *database exercises* section at the end of most chapters. Appendix B describes the donations database on the companion website (giving characteristics of 20,000 individuals together with the amount that they contributed in response to a mailing) that is introduced in Chapter 1 and used in many chapters to illustrate how statistical methods can be used for data mining with Big Data. Appendix C gives detailed solutions to selected parts of problems and database exercises (marked with an asterisk in the text). Appendix D collects all of the statistical tables used throughout the text.

POWERPOINT SLIDES

A complete set of PowerPoint slides, that I developed for my own classes, is available on the companion website.

COMPANION WEBSITE

The companion site http://store.elsevier.com/9780128042502/ includes the PowerPoint presentation slides and Excel files with all quantitative examples and problem data.

INSTRUCTOR'S MANUAL

The instructor's manual is designed to help save time in preparing lectures. A brief discussion of teaching objectives and how to motivate students is provided for each chapter. Also included are detailed solutions to questions, problems, and database exercises, as well as analysis and discussion material for each case. The instructor's manual is available at the instructor website.

ACKNOWLEDGMENTS

Many thanks to all of the reviewers and students who have read and commented on drafts and previous editions of *Practical Business Statistics* over the years. I have been lucky to have dedicated, careful readers at a variety of institutions who were not afraid to say what it would take to meet their needs.

I am fortunate to have been able to work with my parents, Mildred and Armand Siegel, who provided many careful and detailed suggestions for the text.

Very special thanks go to Graham Nisbet, Susan Ikeda, Paul Prasad Chandramohan, and Mark Rogers, who have been very helpful and encouraging with the development and production of this edition. Warm thanks go to Michael Antonucci, who initiated this whole project when he stopped by my office to talk about computers and see what I was up to and encourage me to write it all down. I am also grateful to those who were involved with previous editions, including Lauren Schultz Yuhasz, Lisa Lamenzo, Gavin Becker, Jeff Freeland, Scott Isenberg, Christina Sanders, Catherine Schultz, Richard T. Hercher, Carol Rose, Gail Korosa, Ann Granacki, Colleen Tuscher, Adam Rooke, Ted Tsukahara, and Margaret Haywood. It is a big job producing a work like this, and I was lucky to have people with so much knowledge, dedication, and organizational skill.

Thanks also go out to David Auer, Eric Russell, Dayton Robinson, Eric J. Bean, Michael R. Fancher, Susan Stapleton, Sara S. Hemphill, Nancy J. Silberg, A. Ronald Hauver, Hirokuni Tamura, John Chiu, June Morita, Brian McMullen, David B. Foster, Pablo Ferrero, Rolf R. Anderson, Gordon Klug, Reed Hunt, E.N. Funk, Rob Gullette, David Hartnett, Mickey Lass, Judyann Morgan, Kimberly V. Orchard, Richard Richings, Mark Roellig, Scott H. Pattison, Thomas J. Virgin, Carl Stork, Gerald Bernstein, and Jeremiah J. Sullivan.

A special mention is given to a distinguished group of colleagues who have provided helpful guidance, including Bruce Barrett, University of Alabama; Brian Goff, Western Kentucky University; Anthony Seraphin, University of Delaware; Abbott Packard, Hawkeye Community College; William Seaver, University of Tennessee, Knoxville; Nicholas Jewell, University of California, Berkeley; Howard Clayton, Auburn University; Giorgio Canarello, California State University, Los Angeles; Lyle Brenner, University of Florida, Gainesville; P.S. Sundararaghavan, University of Toledo; Julien Bramel, Columbia University, Ronald Bremer, Texas Tech University; Stergios Fotopoulos, Washington State University; Michael Ghanen, Webster University; Phillip Musa, Texas Tech University; Thomas Obremski, University of Denver; Darrell Radson, University of Wisconsin, Milwaukee; Terrence Reilly, Babson College; Peter Schuhmann, University of Richmond; Bala Shetty, Texas A&M University; L. Dwight Sneathen Jr., University of Arizona; Ted Tsukahara, St. Mary's College; Edward A. Wasil, American University; Michael Wegmann, Keller Graduate School of Management; Mustafa Yilmaz, Northeastern University; Gary Yoshimoto, St. Cloud State University; Sangit Chatterjee, Northeastern University; Jay Devore, California Polytechnic State University; Burt Holland, Temple University; Winston Lin, State University of New York at Buffalo; Herbert Spirer, University of Connecticut; Donald Westerfield; Webster University; Wayne Winston, Indiana University; Jack Yurkiewicz, Pace University; Betty Thorne, Stetson University; Dennis Petruska, Youngstown State University; H. Karim, West Coast University; Martin Young, University of Michigan; Richard Spinetto, University of Colorado at Boulder; Paul Paschke, Oregon State University; Larry Ammann, University of Texas at Dallas; Donald Marx, University of Alaska; Kevin Ng, University of Ottawa; Rahmat Tavallali, Walsh University; David Auer, Western Washington University; Murray Cote, Texas A&M University; Peter Lakner, New York University; Donald Adolphson, Brigham Young University; and A. Rahulji Parsa, Drake University.

TO THE STUDENT

As you begin this course, you may have some preconceived notions of what statistics is all about. If you have positive notions, please keep them and share them with your classmates. But if you have negative notions, please set them aside and remain open-minded until you have given statistics another chance to prove its value in analyzing business risk and providing insight into piles of numbers.

In some ways, statistics is easier for your generation than for those of the past. Now that computers can do the messy numerical work, you are free to develop a deeper understanding of the concepts and how they can help you compete over the course of your business career.

Make good use of the introductory material so that you will always know why statistics is worth the effort. Focus on examples to help with understanding and motivation. Take advantage of the summary, key words, and other materials at the ends of the chapters. Do not forget about the detailed problem solutions and the glossary at the back when you need a quick reminder! And do not worry. Once you realize how much statistics can help you in business, the things you need to learn will fall into place much more easily.

Why not keep this book as a reference? You will be glad you did when the boss needs you to draft a memo immediately that requires a quick look at some data or a response to an adversary's analysis. With the help of *Practical Business Statistics* on your bookshelf, you will be able to finish early and still go out to dinner. Bon appétit!

ANDREW F. SIEGEL

About the Author

Andrew F. Siegel is Professor, Departments of ISOM (Information Systems and Operations Management) and Finance, at the Foster School of Business, University of Washington, Seattle. He is also Adjunct Professor in the Department of Statistics. He has a Ph.D. in statistics from Stanford University (1977), an M.S. in mathematics from Stanford University (1975), and a B.A. in mathematics and physics summa cum laude with distinction from Boston University (1973). Before settling in Seattle, he held teaching and/or research positions at Harvard University, the University of Wisconsin, the RAND Corporation, the Smithsonian Institution, and Princeton University. He has also been a visiting professor at the University of Burgundy at Dijon, France; at the Sorbonne in Paris; and at HEC Business School near Paris. The very first time he taught statistics in a business school (University of Washington, 1983) he was granted the Professor of the Quarter award by the MBA students. He was named the Grant I. Butterbaugh Professor beginning in 1993; this endowed professorship was created by a highly successful executive in honor of Professor Butterbaugh, a business statistics teacher. (Students: Perhaps you will feel this way about your teacher 20 years from now.) Other honors and awards include Excellence in Teaching Awards 2016, 2015, 2014, 2013, 1986, and 1988; Burlington Northern Foundation Faculty Achievement Awards, 1986 and 1992; Research Associate, Center for the Study of Futures Markets, Columbia University, 1988; Research Opportunities in Auditing Award, Peat Marwick Foundation, 1987; and Phi Beta Kappa, 1973.

He belongs to the American Statistical Association where he has served as Secretary-Treasurer of the Section on Business and Economic Statistics. He has written three other books: *Statistics and Data Analysis: An Introduction* (Second Edition, Wiley, 1996, with Charles J. Morgan), *Counterexamples in Probability and Statistics* (Wadsworth, 1986, with Joseph P. Romano), and *Modern Data Analysis* (Academic Press, 1982, coedited with Robert L. Launer). His articles have appeared in many publications, including the *Journal of the American Statistical Association*, the *Journal of Business*, *Management Science*, the *Journal of Finance*, the *Encyclopedia of Statistical Sciences*, the *American Statistician*, the *Review of Financial Studies*, *Proceedings of the National Academy of Sciences of the United States of America*, the *Journal of Financial and Quantitative Analysis*, *Nature*, the *Journal of Portfolio Management*, the *American Mathematical Monthly*, *Mathematical Finance*, the *Journal of the Royal Statistical Society*, the *Annals of Statistics*, the *Annals of Probability*, the *Society for Industrial and Applied Mathematics Journal on Scientific and Statistical Computing*, *Statistics in Medicine*, *Genomics*, the *Journal of Computational Biology*, *Genome Research*, *Biometrika*, *Journal of Bacteriology*, *Statistical Applications in Genetics and Molecular Biology*, *Discourse Processes*, *Auditing: A Journal of Practice and Theory*, *Contemporary Accounting Research*, the *Journal of Futures Markets*, and the *Journal of Applied Probability*. His work has been translated into Chinese and Russian. He has consulted in a variety of business areas, including election predictions for a major television network, statistical algorithms in speech recognition for a prominent research laboratory, television advertisement testing for an active marketing firm, quality control techniques for a supplier to a large manufacturing company, biotechnology process feasibility and efficiency for a large-scale laboratory, electronics design automation for a Silicon Valley startup, and portfolio diversification analysis for a fund management company.

Part I

Introduction and Descriptive Statistics

Welcome to the world of statistics. This is a world you will want to get comfortable with because you will make better management decisions when you know how to assess the available information and how to ask for additional facts as needed. How else can you expect to manage 12 divisions, 683 products, and 5809 employees? And even for a small business, you will need to understand the larger business environment of potential customers and competitors you operates within. These first five chapters will introduce you to the role of statistics (and data mining with Big Data) in business management (Chapter 1) and to the various types of data sets (Chapter 2). Charts help you see the "big picture" that might otherwise remain obscured in a collection of data. Chapter 3 will show you a good way to see the basic facts about a list of numbers—by looking at a *histogram*. Fundamental summary numbers (such as the average, median, and percentiles) will be explained in Chapter 4. One reason statistical methods are so important is that there is so much *variability* out there that gets in the way of the message in the data. Chapter 5 will show you how to measure the extent of the diversity of your observations, which is also used as the most common measure of business risk.

Chapter 1

Introduction

Defining the Role of Statistics in Business

Chapter Outline

A business executive must constantly make decisions under pressure, often with only incomplete and imperfect information available. Naturally, whatever information is available must be utilized to the fullest extent possible. *Statistical analysis* helps extract information from data and provides an indication of the quality of that information. *Data mining* (of "Big Data") combines statistical methods with computer science and optimization in order to help businesses make the best use of the information contained in large data sets. *Probability* helps you understand risky and random events and provides a way of evaluating the likelihood of various potential outcomes.

Even those who would argue that business decision-making should be based on expert intuition and experience (and therefore should not be overly quantified) must admit that all available relevant information should be considered. Thus, statistical techniques should be viewed as an important part of the decision process, allowing informed strategic decisions to be made that combine executive intuition with a thorough understanding of the facts available. This is a powerful combination.

We begin this chapter with an overview of the competitive advantage provided by a knowledge of statistical methods, followed by some basic facts about statistics and probability and their role in business. Statistical activities can be grouped into five main activities (designing, exploring, modeling, estimating, and hypothesis testing) and one way to clarify statistical thinking is to be able to match the business task at hand with the correct collection of statistical methods. This chapter sets the stage for the rest of the book, which follows up with many important detailed procedures for accomplishing business goals that involve these activities. Next follows an overview of data mining of Big Data (which involves these main activities) and its importance in business. Then we distinguish the field of probability (where, based on assumptions, we reach conclusions about what is likely to happen—a useful exercise in business where nobody knows for sure what will happen) from the field of statistics (where we know from the data what happened, from which we infer conclusions about the system that produced these data) while recognizing that probability and statistics will work well together in future chapters. The chapter concludes with some words of advice on how to integrate statistical thinking with other business viewpoints and activities.

1.1 WHY STATISTICS?

Is knowledge of statistics really necessary to be successful in business? Or is it enough to rely on intuition, experience, and hunches? Let us put it another way: Do you really want to ignore much of the vast potentially useful information out there that comes in the form of data?

Practical Business Statistics, Seventh Edition. http://dx.doi.org/10.1016/B978-0-12-804250-2.00001-8

Why Should You Learn Statistics?

By learning statistics, you acquire the competitive advantage of being comfortable and competent around data and uncertainty. A vast amount of information is contained in data, but this information is often not immediately accessible—statistics helps you extract and understand this information. A great deal of skill goes into creating strategy from knowledge, experience, and intuition. Statistics helps you deal with the knowledge component, especially when this knowledge is in the form of numbers, by answering questions such as, To what extent should you really believe these figures and their implications? and, How should we summarize this mountain of data? By using statistics to acquire knowledge, you will add to the value of your experience and intuition, ultimately resulting in better decision-making.

To highlight the variety of ways in which statistics brings value, here are some quotes about data and management that support the need for data and its analysis in business, where the word "data" is underlined:

More CFOs say they want people who have initiative and can do things like analyze data and present their findings coherently to colleagues.

Source: *"The Plain-Vanilla Accountant" by Kimberly S. Johnson on page B7 of the Wall Street Journal, May 19, 2015.*

The company sells some of the data it gathers from credit- and debit-card transactions to investors and research firms, which mine the information for clues about trends that can move stock prices. … The details are so valuable that some investment firms have paid more than $2 million apiece for an annual subscription.

Source: *"Firm Tracks Cards, Sells Data" by Bradley Hope on page A1 of the Wall Street Journal, August 7, 2015.*

A change in the role of corporate chief marketing officers (CMO) is considered in which the mining and analysis of data on consumer preferences behavior has become the most essential task in managing marketing. … Thanks to an explosion of data from social-media platforms, call centers, transactions, loyalty programs, registries and more, CMOs who want a seat at the table will have to harness customer data and leverage it – or risk being relegated to chief promotions officer.

Source: *"When CMOs learn to love data, they'll be VIPs in the C-suite" by Natalie Zmuda on page 2 of Advertising Age, volume 83 issue 7, February 13, 2012.*

Companies increasingly are relying on number crunching rather than a top merchant's instinct as they try to combat sluggish sales and changing shopper behavior. Driving the trend are big-data tools popularized by online retailers that take the guesswork out of picking goods. … Wal-Mart has started using Google Analytics data this year to pinpoint holiday food, ingredients and recipe searches by state to help guide decisions about what food to stock in each part of the country in coming months. … The data shape what products will get prime space at the end of aisles and Wal-Mart emails that promote deals or recipes.

Source: *"In Retail, Data Elbows Aside Chief Merchants" by Suzanne Kapner on page B7 of the Wall Street Journal, September 23, 2015.*

You will not be able to avoid statistics. These methods are already used routinely throughout the corporate world, and the lower cost of computing is increasing your need to be able to make decisions based on quantitative information.

Is Statistics Difficult?

It is much easier to become an expert *user* of statistics than it is to become an expert statistician trained in all of the fine details, although some attention to details and computations is very helpful. Learning statistics is much easier than it used to be now that you can concentrate on interpreting the results and their meaning, leaving the repetitive number-crunching tasks to computer software. Although a few die-hard purists may bemoan the decline of technical detail in statistics teaching, it is good to see that these details are now in their proper place; life is too short for all human beings to work out the intricate details of techniques such as long division and matrix inversion. Statistics is no more difficult than any other field of study, and some hard work will be helpful to achieve understanding of the general ideas and concepts in order to effectively apply them in your work.

How Does Learning Statistics Increase Your Decision-Making Flexibility?

Knowledge of statistics *enhances* your ability to make good decisions. Statistics is not a rigid, exact science and should not get in the way of your experience and intuition. By learning about data and the basic properties of uncertain events, you will help solidify the information on which your decisions are based, and you will add a new dimension to your intuition. Think of statistical methods as a component of decision-making, but not the whole story. You want to supplement—not replace—business experience, common sense, and intuition.

1.2 WHAT IS STATISTICS?

Statistics is the art and science of collecting and understanding data. Since *data* refers to any kind of recorded information, statistics plays an important role in many human endeavors.

Statistics Looks at the Big Picture

When you have a large, complex assemblage of many small pieces of information, statistics can help you classify and analyze the situation, providing a useful overview and summary of the fundamental features in the data. If you do not yet have the data, then statistics can help you collect them, ensuring that your questions can be answered and that you spend enough (but not too much) effort in the process.

Statistics Does Not Ignore the Individual

If used carefully, statistics pays appropriate attention to all individuals. A complete and careful statistical analysis will summarize the general facts that apply to most individuals and *will also alert you to any exceptions*. If there are special cases in the data that are not adequately summarized in the "big picture," the statistician's job is not yet complete. For example, you may read that in 2014 the average US household size was 2.54 people.[1] Although this is a useful statistic, it does not come close to giving a complete picture of the sizes of all households in the United States. As you will see, statistical methods can easily be used to describe the entire distribution of household sizes.

Example
Data in Management

Data sets are very common in management. Here is a short list of kinds of everyday managerial information that are, in fact, data:

1. Financial statements (and other accounting numbers);
2. Security (stocks, bonds, etc.) prices and volumes and interest rates (and other investment information);
3. Money supply figures (and other government announcements);
4. Sales reports (and other internal company records);
5. Market survey results (and other marketing data);
6. Production quality measures (and other manufacturing records);
7. Human resource productivity records (and other internal databases);
8. Product price and quantity sold for every transaction (and other sales data);
9. Publicity expenditures and results (and other advertising information).

Think about it. Probably much of what you do depends at least indirectly on data. Perhaps someone works for you and advises you on these matters, but you rarely see the actual data. From time to time, you might consider asking to see the "raw data" in order to keep some perspective. Looking at data and asking some questions about them may reveal surprises: You may find out that the quality of the data is not as high as you had thought (you mean that is what we base our forecasts on?), or you may find out the opposite and be reassured. Either way, it is worthwhile.

1. U.S. Census Bureau, accessed at https://www.census.gov/hhes/families/data/households.html on September 23, 2015.

Looking at Data With Pictures and Summaries

What do you see when you look hard at tables of data (eg, the financial pages of the *Wall Street Journal* listing stock price information for many companies)? What does a professional statistician see? The surprising answer to both of these questions often is, not much. You have got to go to work on the numbers—draw pictures of them, compute summaries from them, and so on—before their messages will come through. This is what professional statisticians do; they find this much easier and more rewarding than staring at large lists of numbers for long periods of time. So do not be discouraged if a list of numbers looks to you like, well, a list of numbers.

Statistics in Management

What should a manager know about statistics? Your knowledge should include a broad overview of the basic concepts of statistics, with some (but not necessarily all) details. You should be aware that the world is random and uncertain in many aspects. Furthermore, you should be able to effectively perform two important activities:

1. Understand and use the results of statistical analysis as background information in your work.
2. Play the appropriate leadership role during the course of a statistical study if you are responsible for the actual data collection and/or analysis.

To fulfill these roles, you do not need to be able to perform a complex statistical analysis by yourself. However, some experience with actual statistical analysis is essential for you to obtain the perspective that leads to effective interpretation. Experience with actual analysis will also help you to lead others to sound results and to understand what they are going through. Moreover, there may be times when it will be most convenient for you to do some analysis on your own. Thus, we will concentrate on the ideas and concepts of statistics, reinforcing these with practical examples.

1.3 THE FIVE BASIC ACTIVITIES OF STATISTICS

One important way to maintain perspective when applying statistical methods is to keep in mind which of the five main activities is your *main goal* of the moment. In the beginning stages of a statistical study, either there are not yet any data

or else it has not yet been decided what data to look closely at. The *design* phase will resolve these issues so that useful data will result. Once data are available, an initial inspection is called for, provided by the *exploratory* phase. In the *modeling phase*, a system of assumptions and equations is selected in order to provide a framework for further analysis. A numerical summary of an unknown quantity, based on data, is the result of the *estimation* process. The last of these basic activities is *hypothesis testing*, which uses the data to help you decide what the world is really like in some respect. We will now consider these five activities in turn.

Designing a Plan for Data Collection

Designing a plan for data collection might be called *sample survey design* for a marketing study or *experimental design* for a chemical manufacturing process optimization study. This phase of **designing the study** involves planning the details of data gathering. A careful design can avoid the costs and disappointment of finding out—too late—that the data collected are not adequate to answer the important questions. A good design will also collect just the right amount of data: Enough to be useful, but not so much as to be wasteful. Thus, by planning ahead, you can help ensure that the analysis phase will go smoothly and hold down the cost of the project.

Statistics is particularly useful when you have a large group of people, firms, or other items (the *population*) that you would like to know about but cannot reasonably afford to investigate completely. Instead, to achieve a useful but imperfect understanding of this population, you select a smaller group (the *sample*) consisting of some—but not all—of the items in the population. The process of generalizing from the observed sample to the larger population is known as *statistical inference*. The *random sample* is one of the best ways to select a practical sample, to be studied in detail, from a population that is too large to be examined in its entirety.[2] By selecting randomly, you accomplish two goals:

1. You are guaranteed that the selection process is fair and proceeds without bias; that is, all items have an equal chance of being selected. This assures you that, on average, samples will be representative of the population (although each particular random sample is usually only approximately, and not perfectly, representative).
2. The randomness, introduced in a controlled way during the design phase of the project, will help ensure validity of the statistical inferences drawn later.

2. Details of random sampling will be presented in Chapter 8.

Exploring the Data

As soon as you have a set of data, you will want to check it out. **Exploring the data** involves looking at your data set from many angles, describing it, and summarizing it. In this way you will be able to make sure that the data are really what they are claimed to be and that there are no obvious problems.[3] But good exploration also prepares you for the formal analysis in either of two ways:

1. By verifying that the expected features and relationships actually exist in the data, thereby validating the planned techniques of analysis;
2. By finding some unexpected structure in the data that must be taken into account, thereby suggesting some changes in the planned analysis.

Exploration is the first phase once you have data to look at. It is often not enough to rely on a formal, automated analysis, which can be only as good as the data that go into the computer and which assumes that the data set is "well behaved." Whenever possible, examine the data directly to make sure they look OK; that is, there are no large errors, and the relationships observable in the data are appropriate to the kind of analysis to be performed. This phase can help in (1) editing the data for errors, (2) selecting an appropriate analysis, and (3) validating the statistical techniques that are to be used in further analysis.

Modeling the Data

In statistics, a **model** is a system of assumptions and equations that can generate artificial data similar to the data you are interested in, so that you can work with a few numbers (called *parameters*) that represent the important aspects of the data. A model can be a very effective system within which questions about large-scale properties of the data can be answered.

Having the additional structure of a statistical model can be important for the next two activities of estimation and hypothesis testing. We often try to explore the data before deciding on the model, so that you can discover whatever structure—whether expected or unexpected—is actually in the data. In this way, data exploration can help you with modeling. Often, a model says that

data equals structure plus random noise.

For example, with a data set of 3,258 numbers, a model with a single parameter representing "average additional sales dollars generated per dollar of advertising expense" could help you study advertising effectiveness by adjusting this parameter until the model produces artificial data

3. Data exploration is used throughout the book, where appropriate, and especially in Chapters 3, 4, 11, 12, and 14.

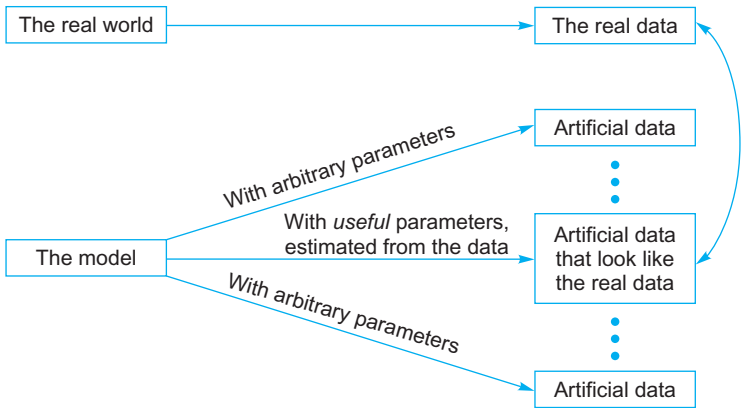

FIG. 1.3.1 A model is a system of assumptions and equations that can generate artificial data. When you carefully choose the parameters of the model, the artificial data (from the model) can be made similar to the real data, and these useful parameters help you understand the real situation.

similar to the real data. Fig. 1.3.1 illustrates how a model, with useful parameters, can be made to match a real data set.

Here are some models that can be useful in analyzing data. Notice that each model generates data with the general approach "data equals structure plus noise," specifying the structure in different ways. In selecting a model, it can be very useful to consider what you have learned by exploring the data.

1. Consider a simple model that generates artificial data consisting of a *single number* plus noise. Chapter 4 (landmark summaries) shows how to extract information about the single number, while Chapter 5 (variability) shows how to describe the noise.
2. Consider a model that generates *pairs* of artificial noisy data values that are related to each other. Chapters 11 and 12 (correlation, regression, and multiple regression) show some useful models for describing the nature and extent of the relationship and the noise.
3. Consider a model that generates a *series* of noisy data values where the next one is related to the previous one. Chapter 14 (time series) presents two systems of models that have been useful in working with business time series data.

Estimating an Unknown Quantity

Estimating an unknown quantity produces the best-educated guess possible based on the available data. We all want (and often need) estimates of things that are just plain impossible to know exactly. Here are some examples of unknowns to be estimated:

1. Next quarter's sales.
2. What the government will do next to our tax rates.
3. How the population of Chicago will react to a new product.
4. How your portfolio of investments will fare next year.
5. The productivity gains of a change in strategy.
6. The defect rate in a manufacturing process.
7. The winners in the next election.
8. The long-term health effects of tablet computer screens.

Statistics can shed light on some of these situations by producing a good, educated guess when reliable data are available. Keep in mind that all statistical estimates are just guesses and are, consequently, often wrong. However, they will serve their purpose when they are close enough to the unknown truth to be useful. If you knew how accurate these estimates were (even approximately), you could decide how much attention to give them as a manager.

Statistical estimation also provides an indication of the amount of uncertainty or error involved in the guess, accounting for the consequences of random selection of a sample from a large population. The *confidence interval* gives probable upper and lower bounds on the unknown quantity being estimated, as if to say, I am not sure exactly what the answer is, but I am quite confident it's between these two numbers.

You should routinely expect to see confidence intervals (and ask for them if you do not) because they show you how reliable an estimated value actually is. For example, there is certainly some information in the forecasting statement that sales next quarter are expected to be

$11.3 million.

However, additional and deeper understanding comes from also being told that you are 95% confident that next quarter's sales will be

between $5.9 million and $16.7 million.

The confidence interval puts the estimate in perspective and helps you avoid the tendency to treat a single number as very precise when, in fact, it might not be precise at all.[4]

Hypothesis Testing

Statistical **hypothesis testing** is the use of data in deciding between two (or more) different possibilities in order to resolve an issue in an ambiguous situation. Hypothesis testing produces a definite decision about which of the possibilities is correct, based on data. The procedure is to collect data that will help decide among the possibilities and to use careful statistical analysis for extra power when the answer is not obvious from just glancing at the data.[5]

Here are some examples of hypotheses that might be tested using data:

1. An Internet advertisement is more effective if placed on the left than on the right of the page.
2. The average New Yorker plans to spend at least $10 on your product next month.
3. You will win tomorrow's election.
4. A new medical treatment is safe and effective.
5. Brand X produces a whiter, brighter wash.
6. The error in a financial statement is smaller than some material amount.
7. It is possible to predict the stock market based on careful analysis of the past.
8. The manufacturing defect rate is below that expected by customers.

Note that each hypothesis makes a definite statement, and it may be either true or false. The result of a statistical hypothesis test is the conclusion that either the data are reasonably consistent with a hypothesis or they are "significantly different."

Often, statistical methods are used to decide whether you can rule out "pure randomness" as a possibility. For example, if a poll of 300 people shows that 53% plan to vote for you tomorrow, can you conclude that the election will go in your favor? Although many issues are involved here, we will (for the moment) ignore details, such as the (real) possibility that some people will change their minds between now and tomorrow, and instead concentrate only on the element of randomness (due to the fact that you cannot call and ask every voter's preference). In this example, a careful analysis would reveal that it is a real possibility that less than 50% of voters prefer you and that the 53% observed is within the range of the expected random sampling variation. For executives, hypothesis testing often plays the valuable role of a filter to help you decide which data items are worth your managerial attention so that this attention is not wasted on random artifacts of statistical noise.

Example

Statistical Quality Control

Your manufacturing processes are not perfect (nobody's are), and every now and then a product has to be reworked or tossed out. Thank goodness for your inspection team, which keeps these bad pieces from reaching the public. Meanwhile, however, you are losing lots of money manufacturing, inspecting, fixing, and disposing of these problems. This is why so many firms have begun using statistical quality control.

To simplify the situation, consider your assembly line to be *in control* if it produces similar results over time that are within the required specifications. Otherwise, your line will be considered to be *out of control*. Statistical methods help you monitor the production process so that you can save money in three ways: (1) Keep the monitoring costs down, (2) detect problems quickly so that waste is minimized, and (3) whenever possible, do not spend time fixing it if it is not broken. Following is an outline of how the five basic activities of statistics apply to this situation.

During the design phase, you have to decide *what* to measure and *how often* to measure it. You might decide to select a random sample of five products to represent every batch of 500 produced. For each one sampled, you might have someone (or something) measure its length and width as well as inspect it visually for any obvious flaws. The result of the design phase is a plan for the early detection of problems. The plan must work in *real time* so that problems are discovered immediately, not next week.

Data exploration is accomplished by plotting the measured data on *quality-control charts* and looking for patterns that suggest trouble. By spotting trends in the data, you may even be able to anticipate and fix a problem before any production is lost!

In the modeling phase, you might choose a standard statistical model, asserting that the observed measurements fluctuate randomly about a long-term average. Such a model then allows you to estimate both the long-term average and the amount of randomness, and then to test whether these values are acceptable.

Statistical estimation can provide management with useful answers to questions about how the production process is going. You might assign a higher grade of quality to the production when it is well controlled within precise limits; such high-grade items command a higher price. Estimates of the quality grade of the current production will be needed to meet current orders, and forecasting of future quality grades will help with strategic planning and pricing decisions.

4. Details of confidence intervals will be presented in Chapter 9 and used in Chapters 9–15.

5. Details of hypothesis testing will be presented in Chapter 10 and used in Chapters 10–18.

Example—cont'd

Statistical hypothesis testing can be used to answer the important question: Is this process in control, or has it gone out of control? Because a production process can be large, long, and complicated, you cannot always tell just by looking at a few machines. By making the best use of the statistical information in your data, you hope to achieve two goals. First, you want to detect when the system has gone out of control even before the quality has become unacceptable. Second, you want to minimize the "false alarm" rate so that you are not always spending time and money trying to fix a process that is really still in control.

Example

A New Product Launch

Deciding whether or not to launch a new product is one of the most important decisions a company makes, and many different kinds of information can be helpful along the way. Much of this information comes from statistical studies. For example, a marketing study of the target consumer group could be used to estimate how many people would buy the product at each of several different prices. Historical production-cost data for similar items could be used to assess how much it would cost to manufacture. Analysis of past product launches, both successful and unsuccessful, could provide guidance by indicating what has worked (and failed) in the past. A look at statistical profiles of national and international firms with similar products will help you size up the nature of possible competition. Individual advertisements could be tested on a sample of viewers to assess consumer reaction before spending large amounts on a few selected advertisements.

The five basic activities of statistics show up in many ways. Because the population of consumers is too large to be examined completely, you could *design* a study, choosing a sample to represent the population (eg, to look at consumer product purchase decisions, or for reactions to specific advertisements). Data *exploration* could be used throughout, wherever there are data to be explored, in order to learn about the situation (eg, are there separate groups of customers, suggesting market segmentation?) and as a routine check before other statistical procedures are used. A variety of statistical *models* could be chosen, adapted to specific tasks. One model might include parameters that relate consumer characteristics to their likelihood of purchase, while another model might help in forecasting future economic conditions at the projected time of the launch. Many *estimates* would be computed, for example, indicating the potential size of the market, the likely initial purchase rate, and the cost of production. Finally, various *hypothesis tests* could be used, for example, to tell whether there is sufficient consumer interest to justify going ahead with the project or to decide whether one advertisement is measurably better (instead of just randomly better) than another in terms of consumer reaction.

1.4 DATA MINING AND BIG DATA

Most companies routinely collect data—at the cash register for each purchase, on the factory floor from each step of production, or on the Internet from each visit to its website—resulting in huge databases containing potentially useful information about how to increase sales, how to improve production, or how to turn mouse clicks into purchases. **Data mining** is a collection of methods for obtaining useful knowledge by analyzing large amounts of data (Big Data) often by searching for hidden patterns. Once a business has collected information for some purpose, it would be wasteful to leave it unexplored when it might be useful in many other ways. The goal of data mining is to obtain value from these vast stores of data, in order to improve the company with higher sales, lower costs, and better products. Here are just a few of the many areas of business in which data mining can be helpful:

1. *Marketing and sales*: Companies have lots of information about past contacts with potential customers and their results. These data can be mined for guidance on how (and when) to better reach customers in the future. One example is the difficult decision of when a store should reduce prices: Reduce too soon and you lose money (on items that might have been sold for more); reduce too late and you may be stuck (with items no longer in season). As reported in the *Wall Street Journal*:

   > *A big challenge: trying to outfox customers who have been more willing to wait and wait for a bargain.…The stores analyze historical sales data to pinpoint just how long to hold out before they need to cut a price—and by just how much.… The technology, still fairly new and untested, requires detailed and accurate sales data to work well.*[6]

   > *… retailers need real-time data to decide when to put something into their storefront, when to discount it and when to take it away to make space for new items.*[7]

   Another example is the supermarket affinity card, allowing the company to collect data on every purchase, while knowing your mailing address. This could allow personalized coupon books to be sent, for example, if no peanut butter had been purchased for 2 months by a customer who usually buys some each month.

6. A. Merrick, "Priced to Move: Retailers Try to Get Leg Up on Markdowns with New Software," *The Wall Street Journal*, August 7, 2001, p. A1.
7. K. Gordon, "Fashion Industry Meets Big Data," *The Wall Street Journal*, September 9, 2013, p. B7.

2. *Finance*: Mining of financial data can be useful in forming and evaluating investment strategies and in hedging (reducing) risk. In the stock markets alone, there are many companies: About 3,292 listed on the New York Stock Exchange and about 3,100 companies listed on the NASDAQ Stock Market.[8] Historical information on price and volume (number of shares traded) is easily available (eg, at http://finance.yahoo.com) to anyone interested in exploring investment strategies. Statistical methods, such as hypothesis testing, are helpful as part of data mining to distinguish random behavior from systematic behavior because stocks that performed well last year will not necessarily perform well next year. Imagine that you toss 100 coins six times each and then carefully choose the one that came up "heads" all six times—this coin is not as special as it might seem!
3. *Product design*: What particular combinations of features are customers ordering in larger-than-expected quantities? The answers could help you create products to appeal to a group of potential customers who would not take the trouble to place special orders.
4. *Production*: Imagine a factory running 24/7 with thousands of partially completed units, each with its bar code, being carefully tracked by the computer system, with efficiency and quality being recorded as well. This is a tremendous source of information that can tell you about the kinds of situations that cause trouble (such as finding a machine that needs adjustment by noticing clusters of units that do not work) or the kinds of situations that lead to extra-fast production of the highest quality.
5. *Fraud detection*: Fraud can affect many areas of business, including consumer finance, insurance, and networks (including telephone and the Internet). One of the best methods of protection involves mining data to distinguish between ordinary and fraudulent patterns of usage, then using the results to classify new transactions, and looking carefully at suspicious new occurrences to decide whether or not fraud is actually involved. I once received a telephone call from my credit card company asking me to verify recent transactions—identified by its statistical analysis—that departed from my typical pattern of spending. In particular, PayPal, with its digital payment systems, pays close attention to this issue. Consider[9]:

Several kinds of algorithms analyze thousands of data points in real-time, such as IP address, buying history, recent activity at the merchant's site or at PayPal's site and information stored in cookies. Results are compared with external data from identity authentication providers. Each transaction is scored for likely fraud, with suspicious activity flagged for further automated and human scrutiny.

Data mining is a large task that involves combining resources from many fields. Here is how statistics, computer science, and optimization are used in data mining:

- *Statistics*: All of the five basic activities of statistics are involved: A design for collecting the data, exploring for patterns, a modeling framework, estimation of features, and hypothesis testing to assess significance of patterns as a "reality check" on the results. Nearly every method in the rest of this book has the potential to be useful in data mining, depending on the database and the needs of the company. Some specialized statistical methods are particularly useful, including *classification analysis* (also called *discriminant analysis*) to assign a new case to a category (such as "likely purchaser" or "fraudulent"), *cluster analysis* to identify homogeneous groups of individuals, and *prediction analysis* (also called *regression analysis*).
- *Computer science*: Efficient algorithms (computer instructions) are needed for collecting, maintaining, organizing, and analyzing data because simpler methods would be too slow. Creative methods involving *artificial intelligence* are useful, including *statistical machine learning* techniques for prediction analysis such as *neural networks* and *boosting*, to learn from the data by identifying useful patterns automatically. Some of these methods from computer science are closely related to statistical prediction and regression analysis.
- *Optimization*: These minimization and maximization methods help you achieve a numeric goal, which might be very specific such as maximizing profits, lowering production cost, finding new customers, developing profitable new products, or increasing sales volume. Alternatively, the goal might be more vague such as obtaining a better understanding of the different types of customers you serve, characterizing the differences in production quality that occur under different circumstances, or identifying relationships that occur more or less consistently throughout the data. Optimization is often accomplished by *adjusting the parameters of a model* until the objective is achieved.

8. Information accessed at http://www.nasdaq.com/screening/companies-by-industry.aspx?exchange=NASDAQ on September 23, 2015.
9. K.S. Nash, "PayPal Fights Fraud with Machine Learning and 'Human Detectives'," *The Wall Street Journal*, dated August 25, 2015, accessed at http://blogs.wsj.com/cio/2015/08/25/paypal-fights-fraud-with-machine-learning-and-human-detectives/ September 23, 2015.

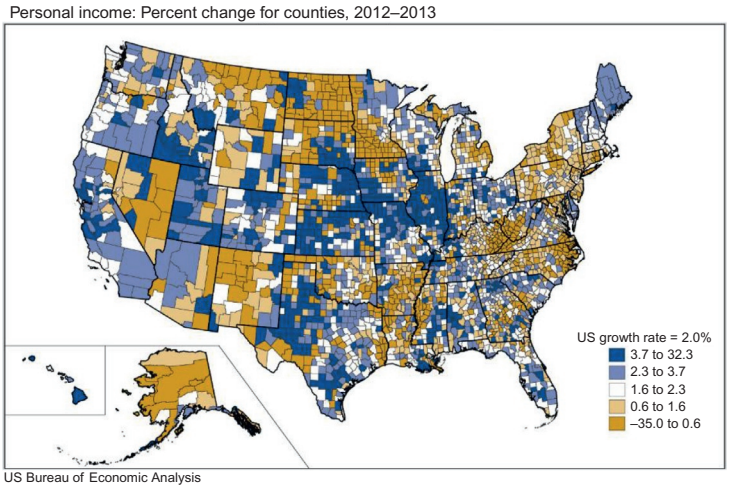

FIG. 1.4.1 Change in personal income displayed county by county as estimated by the US BEA as part of its mission to provide relevant and accurate economic data. Availability of free government data and estimates like this provides opportunities for data mining to help businesses better understand their customers and where they live. *(Source: US Bureau of Economic Analysis, accessed at http://www.bea.gov/newsreleases/regional/lapi/lapi_newsrelease.htm on September 21, 2015.)*

Example
Mining US Neighborhood Data for Potential Customers

Ideally, when deciding where to locate a new store, restaurant, or factory, or where your company should send its catalog, you would want to look everywhere in the whole country (perhaps even beyond) before deciding. A tremendous amount of information is collected, both by the government and by private companies, on the characteristics of neighborhoods across the United States. The US Bureau of Economic Analysis (BEA) a government agency, does much more than just calculate and publish information about GDP (gross domestic product, a measure of production), consumption, investment, exports, imports, income, and savings. In particular, the BEA also produces estimates of personal income for over 3,000 counties as shown in Fig. 1.4.1, which combines information about many types of income (eg, working, owing a business, renting, investing) from many sources. Their estimates are described as follows[10]:

The state and county personal income and employment estimates are based primarily on administrative records data. In addition, some survey and census data are used. The administrative records data are a byproduct of the administration of various federal and state government social insurance programs and tax codes. They may originate either from the recipients of the income or from the payer of the income.

Private companies also collect and analyze detailed information on the characteristics of US neighborhoods. One such company is Experian, which maintains the Mosaic system. Considerable data mining and statistical methods went into classifying consumers by developing 71 segments within 19 groups using over 300 data factors, and much more statistical work goes into providing detailed current information on specific neighborhoods to businesses to help them find customers. Fig. 1.4.2 shows the some of the system's segments, which were developed in part by analyzing data as follows[11]:

… The clustering techniques utilize a multidimensional approach to ensure that all individual, household, and geographic characteristics that will influence consumer behavior are considered and explained. Extensive verification and testing is undertaken to assure that performance is optimized and segments reflect real-world consumer perspectives.

10. Accessed at http://www.bea.gov/regional/pdf/lapi2013.pdf on September 21, 2015. Some the programs, taxes, and agencies that contribute statistical information to the BEA's estimates include state unemployment insurance programs, Bureau of Labor Statistics at U.S. Department of Labor, state Medicaid programs, Centers for Medicare and Medicaid Services, U.S. Department of Health and Human Services, Social Security Administration, U.S. Department of Veterans Affairs, state and federal income tax codes, Internal Revenue Service at U.S. Department of the Treasury, and Bureau of the Census at U.S. Department of Commerce.
11. Accessed at http://www.appliedgeographic.com/AGS_mosaic_2012/Mosaic_Methodology.pdf on September 23, 2015.

Example
Mining Data to Identify People Who Will Donate to a Good Cause

Many people send money to charity in response to requests received in the mail, but many more do not respond—and sending letters to these nonresponders is costly. If you worked for a charitable organization, you would want to be able to predict the likelihood of

(*Continued*)

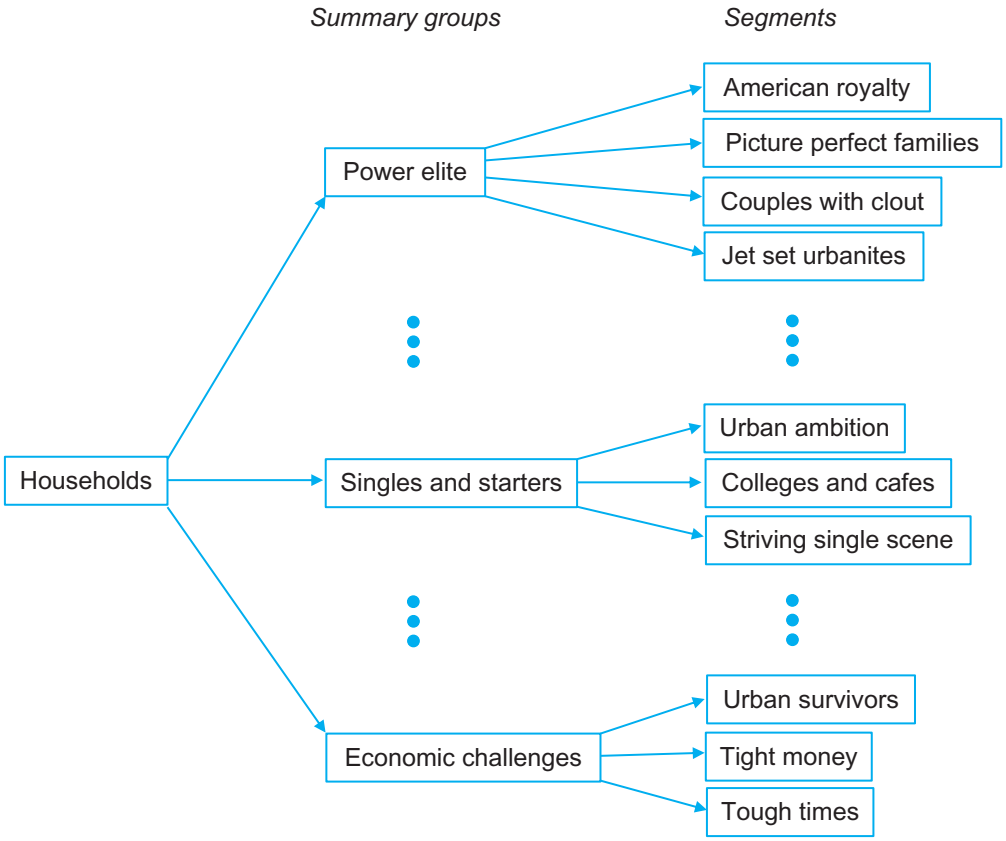

FIG. 1.4.2 Some results of data mining to identify clusters of households that tend to be similar across many observable characteristics, as determined by the Mosaic system from Experian for understanding customers and markets. There are two levels of clusters. Based on data, each household sampled from a neighborhood can be classified into one of 19 summary groups (the top level of clusters) of which three are shown, and further classified into one of the 71 detailed segments (the next level of clusters) for its summary group. There are many business uses for systems that can help find customers, including where to locate a store and where to send mailings. *(Source: Based on information accessed at https://www.experian.com/assets/marketing-services/brochures/mosaic-brochure.pdf on September 23, 2015.)*

Example—cont'd

donation and the likely amount of the donation ahead of time—before sending a letter—to help you decide where and when to send a request for money. Managers of nonprofit companies (such as charities) need to use many of the same techniques as those of for-profit companies, and data-mining methods can be very helpful to a manager of any company hoping to make better use of data collected on the results of past mailings (and Web screens) in order to help plan for the future.

A difficult decision is how often to keep sending requests to people who have responded in the past, but not recently. Some of them will become active donors again—but which ones? Table 1.4.1 shows part of a database that gives information on 20,000 such individuals at the time of a mailing, together with the amount (if any, in the first column) that each one gave as a result of that mailing.[12] The columns in the database are defined in Table 1.4.2. We will revisit this database in future chapters—from description, through summaries, statistical inference, and prediction—to show how many of the various statistical techniques can be used to help with data mining. One quick discovery is shown in Fig. 1.4.3: Apparently the more gifts given over the previous 2 years (from the column headed "Recent Gifts"), the greater the chances that the person gave a gift in response to this mailing.

12. This database was adapted from a large data set originally used in The Second International Knowledge Discovery and Data Mining Tools Competition and is available as part of the UCI Knowledge Discovery in Databases Archive; Hettich, S. and Bay, S.D., 1999, The UCI KDD Archive http://kdd.ics.uci.edu, Irvine, CA, University of California, Department of Information and Computer Science, now maintained as part of the UCI Machine Learning Archive at http://archive.ics.uci.edu/ml/.

TABLE 1.4.1 Charitable Donations Mailing Database[a]

Donation ($)	Lifetime ($)	Gifts	Years Since First	Years Since Last	Average Gift ($)	Major Donor	Promos	Recent Gifts	Age	Home Phone	PC Owner	Catalog Shopper	Per Capita Income	Median Household Income	Professional (%)	Technical (%)	Sales (%)	Clerical (%)	Farmers (%)	Self-Employed (%)	Cars (%)	Owner Occupied (%)	Age 55–59 (%)	Age 60–64 (%)	School
0.00	81.00	15	6.4	1.2	5.40	0	58	3		0	0	0	16,838	30,500	12	7	17	22	1	2	16	41	4	5	14.0
15.00	15.00	1	1.2	1.2	15.00	0	13	1	33	1	0	1	17,728	33,000	11	1	14	16	1	6	8	90	7	11	12.0
0.00	15.00	1	1.8	1.8	15.00	0	16	1		1	0	0	6,094	9,300	3	0	5	32	0	0	3	12	6	3	12.0
0.00	25.00	2	3.5	1.3	12.50	0	26	1	55	0	0	0	16,119	50,200	4	7	16	19	6	21	52	79	3	2	12.3
0.00	20.00	1	1.3	1.3	20.00	0	12	1	71	1	0	0	11,236	24,700	7	3	7	15	2	5	22	78	6	6	12.0
0.00	68.00	6	7.0	1.6	11.33	0	38	2	42	0	0	0	13,454	40,400	15	2	7	4	14	17	26	67	6	5	12.0
0.00	110.00	11	10.2	1.4	10.00	0	38	2	75	1	0	0	8,655	17,000	8	3	5	12	15	15	21	82	8	5	12.0
0.00	174.00	26	10.4	1.5	6.69	0	72	3		0	0	0	6,461	13,800	7	4	9	12	1	4	12	57	6	6	12.0
0.00	20.00	1	1.8	1.8	20.00	0	15	1	67	1	0	0	12,338	37,400	11	2	16	18	3	3	22	90	10	9	12.0
14.00	95.00	7	6.1	1.3	13.57	0	56	2	61	0	0	0	10,766	20,300	13	4	11	8	2	7	20	67	7	7	12.0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
0.00	25.00	2	1.5	1.1	12.50	0	18	2		0	0	1	9,989	23,400	14	2	9	10	0	7	20	73	7	6	12.0
0.00	30.00	2	2.2	1.4	15.00	0	19	1	74	1	0	0	11,691	27,800	4	1	8	14	0	2	10	65	6	8	12.0
0.00	471.00	22	10.6	1.5	21.41	0	83	1	87	0	0	0	20,648	34,000	13	4	20	20	0	2	5	46	8	9	12.4
0.00	33.00	3	6.1	1.2	11.00	0	31	1	42	1	0	0	12,410	21,900	9	3	12	20	0	9	13	49	5	8	12.0
0.00	94.00	10	1.1	0.3	9.40	0	42	1	51	0	0	0	14,436	41,300	15	7	9	15	1	9	29	85	6	5	13.2
0.00	47.00	8	3.4	1.0	5.88	0	24	4	38	0	1	0	17,689	31,800	11	3	17	21	0	6	12	16	2	3	14.0
0.00	125.00	7	5.2	1.2	17.86	0	49	3	58	0	1	0	26,435	43,300	15	1	5	9	0	3	16	89	5	24	14.0
0.00	109.50	16	10.6	1.3	6.84	0	68	4	67	0	0	0	17,904	44,800	8	3	1	20	4	15	26	88	6	5	12.0
0.00	112.00	11	10.2	1.6	10.18	0	66	2	82	0	0	0	11,840	28,200	13	4	12	14	2	6	13	77	5	5	12.0
0.00	243.00	15	10.1	1.2	16.20	0	67	2	67	0	0	0	17,755	40,100	10	3	13	24	2	7	24	41	2	4	14.0

[a]*The first column shows how much each person gave as a result of this mailing, while the other columns show information that was available before the mailing was sent. Data mining can use this information to statistically predict the mailing result, giving us useful information about characteristics that are linked to the likelihood and amount of donations.*

TABLE 1.4.2 Definitions for the Variables in the Donations Database[a]

Name of Variable	Description
Donation	Donation amount in dollars in response to this mailing
Lifetime	Donation lifetime total before this mailing
Gifts	Number of lifetime gifts before this mailing
Years Since First	Years since first gift
Years Since Last	Years since most recent gift before this mailing
Average Gift	Average of gifts before this mailing
Major Donor	Major donor indicator
Promos	Number of promotions received before this mailing
Recent Gifts	Number of gifts in past 2 years
Age	Age in years
Home Phone	Published home phone number indicator
PC Owner	Home PC owner indicator
Catalog Shopper	Shop by catalog indicator
Per Capita Income	Per capita neighborhood income
Median Household Income	Median household neighborhood income
Professional	Percent professional in neighborhood
Technical	Percent technical in neighborhood
Sales	Percent sales in neighborhood
Clerical	Percent clerical in neighborhood
Farmers	Percent farmers in neighborhood
Self-Employed	Percent self-employed in neighborhood
Cars	Percent households with 3+ vehicles
Owner Occupied	Percent owner-occupied housing units in neighborhood
Age 55–59	Percent adults age 55–59 in neighborhood
Age 60–64	Percent adults age 60–64 in neighborhood
School	Median years in school completed by adults in neighborhood

[a]*The first group of variables represents information about the person who received the mailing. For example, the second variable, "Lifetime," shows the total dollar amount of all previous gifts by this person, and variable 12 "PC Owner" is 1 if he or she owns a PC and is 0 otherwise. The remaining variables represent information about the person's neighborhood, beginning with column 14 "Per Capita Income" and continuing through all of the percentages to the last column.*

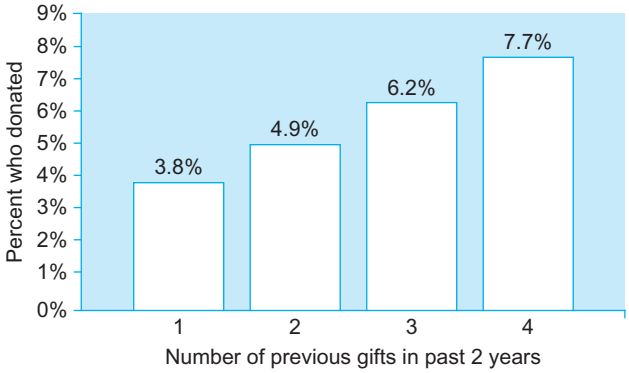

FIG. 1.4.3 A result of data mining of the donations database of 20,000 people. The more gifts given over the previous 2 years (from the database column headed "Recent Gifts"), the greater the chances that the person gave a gift in response to this mailing. For example, out of the 9,924 who gave just one previous gift, 381 (or 3.8%) gave a gift. Out of the 2,486 who gave four previous gifts, 192 (for a larger percentage of 7.7%) donated.

1.5 WHAT IS PROBABILITY?

Probability is a *what if* tool for understanding risk and uncertainty. **Probability** shows you the likelihood, or chances, for each of the various potential future events that might occur, based on a set of assumptions about how the world works. For example, you might assume that you know basically how the world works (ie, all of the details of the process that will produce success or failure or payoffs in between). Probabilities of various outcomes would then be computed for each of several strategies to indicate how successful each strategy would be.

You might learn, for example, that an international project has only an 8% chance of success (ie, the probability of success is 0.08), but if you assume that the government can keep inflation low, then the chance of success rises to 35%—still very risky, but a much better situation than the 8% chance. Probability will not tell you whether to invest in the project, but it will help you keep your eyes open to the realities of the situation.

Here are additional examples of situations where finding the appropriate answer requires computing or estimating a probability number:

1. Given the nature of an investment portfolio and a set of assumptions that describe how financial markets work, what are the chances that you will profit over a 1-year horizon? Over a 10-year horizon?
2. What are the chances of rain tomorrow? What are the chances that next winter will be cold enough so that your heating-oil business will make a profit?

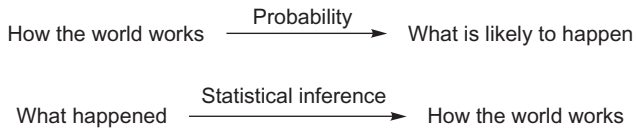

FIG. 1.5.1 Probability and statistics take you in opposite directions. If you make assumptions about how the world works, then probability can help you figure out how likely various outcomes are and thus help you understand what is likely to happen. If you have data that tell you something about what has happened, then statistics can help you move from this particular data set to a more general understanding of how things work.

3. What are the chances that a foreign country (where you have a manufacturing plant) will become involved in civil war over the next 2 years?
4. What are the chances that the smartphone your company just shipped will fail during the first 3 months?
5. What are the chances that the college student you just interviewed for a job will become a valued employee over the coming months?

Probability is the inverse of statistics. Whereas statistics helps you go from observed data to generalizations about how the world works, probability goes the other direction: If you assume you know how the world works, then you can figure out what kinds of data you are likely to see and the likelihood for each. Fig. 1.5.1 shows this inverse relation.

Probability also works together with statistics by providing a solid foundation for statistical inference. When there is uncertainty, you cannot know exactly what will happen, and there is some chance of error. Using probability as part of hypothesis testing, you will learn ways to control a decision's error rate so that it is, say, less than 5% or less than 1% of the time.

1.6 GENERAL ADVICE

Statistical results should be explainable in a straightforward way (even though their theory may be much more complicated), and statistical methods should be used together with (and not replace) expert knowledge in subject areas such as economics and marketing. Here are some general words of advice:

1. Trust your judgment; common sense counts—do not be too quick to change course based on one new data set.
2. Maintain a healthy skepticism—ask for convincing evidence before agreeing with others' assertions.
3. Do not be snowed by a seemingly ingenious statistical analysis; it may well rely on unrealistic and inappropriate assumptions.

Because of the vast flexibility available to the analyst in each phase of a study, one of the most important factors to consider in evaluating the results of a statistical study is: *Who funded it*? Remember that the analyst made many choices along the way—in defining the problems, designing the plan to select the data, choosing a framework or model for analysis, and interpreting the results.

1.7 END-OF-CHAPTER MATERIALS

Summary

Statistics is the art and science of collecting and understanding data. Statistical techniques should be viewed as an important part of the decision process, allowing informed strategic decisions to be made that combine intuition and (nonstatistical) expertise with a thorough (statistical) understanding of the facts available. Use of statistics is becoming increasingly important in maintaining a competitive edge.

The five basic activities of statistics are as follows:

1. **Designing the study** involves planning the details of data gathering, perhaps using a random sample from a larger population.
2. **Exploring the data** involves looking at your data set from many angles, describing it, and summarizing it. This helps you make sure that the planned analysis is appropriate and allows you to modify the analysis if necessary.
3. **Modeling the data** involves choosing a system of assumptions and equations that behaves like the data you are interested in, so that you can work with a few numbers (called *parameters*) that represent the important aspects of the data. A model can be a very effective system within which questions about large-scale properties of the data can be answered. Often, a model has the form "data equals structure plus noise."
4. **Estimating an unknown quantity** produces the best educated guess possible based on the available data. You will also want to have some indication of the size of the error involved when you use this estimated value in place of the (unknown) actual value.
5. **Statistical hypothesis testing** uses data to decide between two (or more) different possibilities in order to resolve an issue in an ambiguous situation. This is often done to see if some apparently interesting feature of the data is really there ("significant") as opposed to being an uninteresting artifact of "pure randomness."

Data mining is a collection of methods for obtaining useful knowledge by analyzing large amounts of data ("Big Data") often by searching for hidden patterns. It would be wasteful to leave this information unexplored, after having been collected for some purpose, when it could be useful in many other ways. The goal of data mining is to obtain value from these vast stores of data, in order to improve the company

with higher sales, lower costs, and better products. Data mining uses all five activities of statistics, plus computer science and optimization.

Probability shows you the likelihood, or chances, for each of the various potential future events that might occur, based on a set of assumptions about how the world works. Probability is the inverse of statistics: Probability tells you what the data will be like when you know how the world is, whereas statistics helps you figure out what the world is like after you have seen some data that it generated.

Statistics works best when you combine it with your own expert judgment and common sense. When statistical results go against your intuition, be prepared to work hard to find the reason why: The statistical analysis may well be incorrect due to wrong assumptions, or your intuition may be wrong because it was not based on facts.

Keywords

Data mining, *9*
Designing the study, *6*
Estimating an unknown quantity, *7*
Exploring the data, *6*
Hypothesis testing, *8*
Model, *6*
Probability, *14*
Statistics, *4*

Questions

1. Why is it worth the effort to learn about statistics?
   a. Please answer for management in general.
   b. Please answer for one particular area of business of special interest to you.
2. Choose a business firm, and list the ways in which statistical analysis could be used in decision-making activities within that firm.
3. How should statistical analysis and business experience interact with each other?
4. What is statistics?
5. What is the design phase of a statistical study?
6. Why is random sampling a good method to use for selecting items for study?
7. What can you gain by exploring data in addition to looking at summary results from an automated analysis?
8. What can a statistical model help you accomplish? Which basic activity of statistics can help you choose an appropriate model for your data?
9. Are statistical estimates always correct? If not, what else will you need (in addition to the estimated values) in order to use them effectively?
10. Why is a confidence interval more useful than an estimated value?
11. Give two examples of hypothesis testing situations that a business firm would be interested in.
12. What distinguishes data mining from other statistical methods? What methods, in addition to those of statistics, are often used in data mining?
13. Differentiate between probability and statistics.
14. A consultant has just presented a very complicated statistical analysis, complete with lots of mathematical symbols and equations. The results of this impressive analysis go against your intuition and experience. What should you do?
15. Why is it important to identify the source of funding when evaluating the results of a statistical study?

Problems

Problems marked with an asterisk (*) are solved in the Self-Test in Appendix C.

1. Describe a recent decision you made that depended, in part, on information that came from data. Identify the underlying data set and tell how a deeper understanding of statistics might have helped you use these data more effectively.
2. Name three numerical quantities a firm might be concerned with for which exact values are unavailable. For each quantity, describe an estimate that might be useful. In general terms, how reliable would you expect these estimates to be?
3. Reconsider the three estimates from the previous problem. Are confidence intervals readily available? How might these confidence intervals be useful?
4. List two kinds of routine decisions that you make in which statistical hypothesis testing could play a helpful role.
5. Look through recent material from the *Wall Street Journal*. Identify an article that relies directly or indirectly on statistics. Briefly describe the article (also be sure to give the title, date, and page number, or URL), and attach a copy. Which of the five activities of statistics is represented here most prominently?
6.* Which of the five basic activities of statistics is represented by each of the following situations?
   a. A factory's quality control division is examining detailed quantitative information about recent productivity in order to identify possible trouble spots.
   b. A focus group is discussing the audience that would best be targeted by advertising, with the goal of drawing up and administering a questionnaire to this group.
   c. In order to get the most out of your firm's Internet activity data, it would help to have a framework or structure of equations to allow you to identify and work with the relationships in the data.
   d. A firm is being sued for gender discrimination. Data that show salaries for men and women are presented to the jury to convince them that there is a consistent pattern of discrimination and that such a disparity could not be due to randomness alone.
   e. The size of next quarter's gross national product must be known so that a firm's sales can be forecast. Since it is unavailable at this time, an educated guess is used.
7. Overseas sales dropped sharply last month, and you do not know why. Moreover, you realize that you do not even have the numbers needed in order to tell what the problem is. You call a meeting to discuss how to solve the problem. Which statistical activity is involved at this stage?

8. If your factories produce too much, then you will have to pay to store the extra inventory. If you produce too little, then customers will be turned away and profits will be lost. Therefore, you would like to produce exactly the right amount to avoid these costs to your company. Unfortunately, however, you do not know the correct production level. Which is the main statistical activity required to solve this problem?
9. Before you proceed with the analysis of a large accounting data set that has just been collected, your boss has asked you to take a close look at the data to check for problems and surprises and ensure its basic integrity. Identify the basic statistical activity you are performing.
10. Your company has been collecting detailed data for years on customer contacts, including store purchases, telephone inquiries, and Internet orders, and you would like to systematically use this resource to learn more about your customers and, ultimately, to improve sales. What is the name of the collection of methods that will be most useful to you in this project?
11. Your work group would like to estimate the size of the market for high-quality stereo components in New Orleans but cannot find any reliable data that are readily available. Which basic activity of statistics is involved initially in proceeding with this project?
12. You are wondering whom to interview, how many to interview, and how to process the results so that your questions can be answered at the lowest possible cost. Identify the basic activity of statistics involved here.
13. You have collected and explored the data on Internet information requests. Before continuing on to use the data for estimation and hypothesis testing, you want to develop a framework that identifies meaningful parameters to describe relationships in the data. What basic activity of statistics is involved here?
14. Your firm has been accused of discrimination. Your defense will argue in part that the imbalance is so small that it could have happened at random and that, in fact, no discrimination exists. Which basic activity of statistics is involved?
15. By looking carefully at graphs of data, your marketing department has identified three distinct segments of the marketplace with different needs and price levels. Which basic activity of statistics helped you to obtain this helpful information?
16. You are trying to determine the quality of the latest shipment of construction materials based on careful observation of a sample. Which basic activity of statistics will help you reach your goal?
17. You think that one of the machines may be broken, but you are not sure because even when it is working properly there are a few badly manufactured parts. When you analyze the rate at which defective parts are being produced to decide whether or not there has been an increase in the defect rate, which basic activity of statistics is involved?
18. Your boss has asked you to take a close look at the marketing data that just came in and would like you to report back with your overall impressions of its quality and usefulness. Which main activity of statistics will you be performing?
19. Using data on the characteristics of houses that sold recently in a city of interest, you would like to specify the way in which features such as the size (in square feet) and number of bedrooms relate to the sale price. You are working out an equation that asserts that the sales price is given by a formula that involves the house's characteristics and parameters (such as the dollar value of an additional bedroom) that are estimated from the data. What main activity of statistics are you involved with?
20. The results of the customer survey just arrived as a spreadsheet from the firm that was hired to do the research, and you are eager to understand what can be learned from these responses. Naturally you will be producing charts and summaries in order to obtain an overall impression of this new information (the firm did not provide these …). Which main activity of statistics is most strongly related to your work?
21. While your longer-term goal is to better understand your customer base by seeing how their attitudes correlate with purchasing activity, you realize that (at the moment) the information you would need to do this is not available. Thus your immediate task is to figure out how to survey customers in depth to gather this information. Which of the main activities of statistics is most directly involved in your immediate task? Please give its name and the reason for your choice.

Projects

Find the results of an opinion poll in a newspaper or magazine or on the Internet. Discuss how each of the five basic activities of statistics was applied (if it is clear from the article) or might have been applied to the problem of understanding what people are thinking. Attach a copy of the article to your discussion.

Chapter 2

Data Structures

Classifying the Various Types of Data Sets

Chapter Outline

Data can come to you in several different forms, and it will be useful to have a basic catalog of the different kinds of data so that you can recognize them and use appropriate techniques for each. A **data set** consists of observations on items, typically with the same information being recorded for each item. We define the **elementary units** as the items themselves (eg, companies, people, households, cities, TV sets) in order to distinguish them from the measurement or observation (eg, sales, weight, income, population, size).

This chapter shows that data sets can be classified in five basic ways:

One: By the number of pieces of information (variables) there are for each elementary unit. Univariate data have just one variable, bivariate data have two variables (eg, cost and number produced), and multivariate data have three or more variables.

Two: By the kind of measurement (numbers or categories) recorded in each case. Quantitative data consist of meaningful numbers, while categorical data are categories that might be ordered ("ordinal data") or not ("nominal data").

Three: By whether or not the time sequence of recording is relevant. Time-series data are more complex to analyze than are cross-sectional data due to the way in which measurements change over time.

Four: By whether or not the information was newly created or had previously been created by others for their own purposes. If you (or your firm) control the data-gathering process, the result is called "primary data" while data produced by others is "secondary data."

Five: By whether the data were merely observed (an "observational study") or if some variables were manipulated or controlled (an "experiment"). Advantages of an experiment include the ability to assess what is causing the reaction of interest.

2.1 HOW MANY VARIABLES?

A piece of information recorded for every item (eg, its cost) is called a **variable**. The number of variables (pieces of information) recorded for each item indicates the complexity of the data set and will guide you toward the proper kinds of analyses. Depending on whether one, two, or many variables are present, you have *univariate, bivariate,* or *multivariate* data, respectively.

Univariate Data

Univariate (one-variable) data sets have just one piece of information recorded for each item. Statistical methods are used to summarize the basic properties of this single piece of information, answering such questions as:

1. What is a typical (summary) value?
2. How diverse are these items?
3. Do any individuals or groups require special attention?

Practical Business Statistics, Seventh Edition. http://dx.doi.org/10.1016/B978-0-12-804250-2.00002-X

Here is a table of univariate data, showing the profits of 10 food services companies in the extended Fortune 500 list.

Company	Profits ($ Millions)
McDonald's	4,758
Starbucks	2,068
Yum Brands	1,051
Darden Restaurants	286
Bloomin' Brands	91
Chipotle Mexican Grill	445
Brinker International	154
Cracker Barrel Old Country Store	132
Panera Bread	179
Wendy's	121

Source: Data from http://fortune.com/fortune500/, accessed October 12, 2015.

Here are some additional examples of univariate data sets:

1. The incomes of subjects in a marketing survey. Statistical analysis would reveal the profile (or distribution) of incomes, indicating a typical income level, the extent of variation in incomes, and the percentage of people within any given income range.
2. The number of defects in each TV set in a sample of 50 manufactured this morning. Statistical analysis could be used to keep tabs on quality (estimate) and to see if things are getting out of control (hypothesis testing).
3. The interest rate forecasts of 25 experts. Analysis would reveal, as you might suspect, that the experts do not all agree and (if you check up on them later) that they can all be wrong. Although statistics cannot combine these 25 forecasts into an exact, accurate prediction, it at least enables you to explore the data for the extent of consensus.
4. The colors chosen by members of a focus group. Analysis could be used to help in choosing an agreeable selection for a product line.
5. The bond ratings of the firms in an investment portfolio. Analysis would indicate the risk of the portfolio.

Bivariate Data

Bivariate (two-variable) data sets have exactly two pieces of information recorded for each item. In addition to summarizing each of these two variables separately (each as its own univariate data set), statistical methods would also be used to explore the relationship between the two factors being measured in the following ways:

1. Is there a simple relationship between the two?
2. How strongly are they related?
3. Can you predict one from the other? If so, with what degree of reliability?
4. Do any individuals or groups require special attention?

Here is a table of bivariate data, showing the profits of 10 food services companies in the extended Fortune 500 list, along with their profits as a percentage of stockholder equity.

Company	Profits ($ Millions)	Profits as Percentage of Stockholder Equity (%)
McDonald's	$4,758	37%
Starbucks	2,068	39
Yum Brands	1,051	67
Darden Restaurants	286	13
Bloomin' Brands	91	16
Chipotle Mexican Grill	445	22
Brinker International	154	244
Cracker Barrel Old Country Store	132	25
Panera Bread	179	24
Wendy's	121	7

Source: Data from http://fortune.com/fortune500/, accessed October 12, 2015.

Here are some additional examples of bivariate data sets:

1. The cost of production (first variable) and the number produced (second variable) for each of seven factories (items, or elementary units) producing integrated circuits, for the past quarter. A bivariate statistical analysis would indicate the basic relationship between cost and number produced. In particular, the analysis might identify a *fixed cost* of setting up production facilities and a *variable cost* of producing one extra circuit.[1] An analyst might then look at individual factories to see how efficient each is compared with the others.
2. The price of one share of your firm's common stock (first variable) and the date (second variable), recorded every day for the past 6 months. The relationship between price and time would show you any recent trends in the value of your investment. Whether or not you could then forecast future value is a subject of some controversy (is it an unpredictable "random walk," or are those apparent patterns real?).
3. The purchase or nonpurchase of an item (first variable, recorded as yes/no or as 1/0) and whether an advertisement for the item is recalled (second variable, recorded similarly) by each of 100 people in a shopping mall. Such data (as well as data from more careful studies) help shed light on the effectiveness of advertising: What is the relationship between advertising recall and purchase?

The reason a bivariate data set can tell you about the relationship (between its two variables) is that these variables were each measured on the *same elementary units*.

1. *Variable cost* refers to the cost that varies according to the number of units produced; it is not related to the concept of a *statistical* variable.

You might have sales and profits for each company, with high profits generally associated with high sales numbers for that same firm. If, instead, you had two univariate data sets representing measurements of different elementary units (say a group of Internet retailers, and a different group of natural resources firms) you would not be able to reach conclusions about a relationship.

Multivariate Data

Multivariate (many-variable) data sets have three or more pieces of information recorded for each item. In addition to summarizing each of these variables separately (as a univariate data set), and in addition to looking at the relationship between any two variables (as a bivariate data set), statistical methods would also be used to look at the interrelationships among all the items, addressing the following questions:

1. Is there a simple relationship among them?
2. How strongly are they related?
3. Can you predict one (a "special variable") from the others? With what degree of reliability?
4. Do any individuals or groups require special attention?

Here is a table of multivariate data, showing the profits of 10 food services companies in the extended Fortune 500 list, along with their profits as a percentage of stockholder equity, number of employees, and revenues.

Company	Profits ($ Millions)	Profits as Percentage of Stockholder Equity (%)	Employees	Revenues ($ Millions)
McDonald's	$4,758	37%	420,000	27,441
Starbucks	2,068	39	191,000	16,448
Yum Brands	1,051	67	303,405	13,279
Darden Restaurants	286	13	206,489	8,758
Bloomin' Brands	91	16	100,000	4,443
Chipotle Mexican Grill	445	22	53,090	4,108
Brinker International	154	244	55,586	2,906
Cracker Barrel Old Country Store	132	25	72,000	2,684
Panera Bread	179	24	35,450	2,529
Wendy's	121	7	31,200	2,061

Source: Data from http://fortune.com/fortune500/, accessed October 12, 2015.

Here are some additional examples of multivariate data sets:

1. The growth rate (special variable) and a collection of measures of strategy (the other variables), such as type of equipment, extent of investment, and management style, for each of a number of new entrepreneurial firms. The analysis would give an indication of which combinations have been successful and which have not.
2. Salary (special variable) and gender (recorded as male/female or as 0/1), number of years of experience, job category, and performance record, for each employee. Situations such as this come up in litigation about whether women are discriminated against by being paid less than men on the average. A key question, which a multivariate analysis can help answer, is, "Can this discrepancy be explained by factors other than gender?" Statistical methods can remove the effects of these other factors and then measure the average salary differential between a man and a woman who are equal in all other respects.
3. The price of a house (special variable) and a collection of variables that contribute to the value of real estate, such as lot size, square footage, number of rooms, presence or absence of swimming pool, and age of house, for each of a collection of houses in a neighborhood. Results of the analysis would give a picture of how real estate is valued in this neighborhood. The analysis might be used as part of an appraisal to determine fair market value of a house in that neighborhood, or it might be used by builders to decide which combination of features will best serve to enhance the value of a new home.

2.2 QUANTITATIVE DATA: NUMBERS

Meaningful numbers are numbers that directly represent the measured or observed *amount* of some characteristic or quality of the elementary units, as the result of an observation of a variable. Meaningful numbers include, for example, dollar amounts, counts, sizes, numbers of employees, and miles per gallon. They *exclude* numbers that are merely used to code for or keep track of something else, such as football uniform numbers or transaction codings like, 1 = buy stock, 2 = sell stock, 3 = buy bond, 4 = sell bond.

If the data for a variable comes to you as meaningful numbers, then you have **quantitative** data (ie, they represent quantities). With quantitative data, you can do all of the usual number-crunching tasks, such as finding the average (see Chapter 4) and measuring the variability (see Chapter 5). It is straightforward to compute directly with numerical data. There are two kinds of quantitative data, *discrete* and *continuous*, depending on the values potentially observable.

Discrete Quantitative Data

A **discrete** variable can assume values only from a list of specific numbers.[2] For example, the number of children

2. Note the difference between a *discrete* variable (as defined here) and a *discreet* variable, which would be much more careful and quiet about its activities.

in a household is a discrete variable. Since the possible values can be listed, it is relatively simple to work with discrete data sets. Here are some examples of discrete variables:

1. The number of network outages in a factory in the past 24 hours.
2. The number of contracts, out of the 18 for which you submitted bids that were awarded.
3. The number of foreign tankers docked at a certain port today.
4. The gender of an employee, if this is recorded using the number 0 or 1.

Continuous Quantitative Data

We will consider any numerical variable that is not discrete to be **continuous**.[3] This word is used because the possible values form a "continuum," such as the set of all positive numbers, all numbers, or all values between 0% and 100%. For example, the actual weight of a candy bar marked "net weight 1.7 ounces" is a continuous random variable; the actual weight might be 1.70235 or 1.69481 ounces instead of exactly 1.7. If you are not yet thinking statistically, you might have assumed that the actual weight was 1.7 ounces exactly; in fact, when you measure in the real world, there are invariably small (and sometimes large) deviations from expected values.

Here are some examples of continuous variables:

1. The price of an ounce of gold, in dollars, right now. You might think that this value is discrete (and you would be technically correct, since a number such as $1,155.90 is part of a list of discrete numbers of pennies: 0.00, 0.01, 0.02, …). However, try to view such cases as examples of continuous data because the discrete divisions are so small as to be unimportant to the analysis. If gold ever began trading at a few cents per ounce, it would become important to view it as a case of discrete data; however, it is more likely that the price would be quoted in thousandths of a cent at that point, again essentially a continuous quantity.
2. Investment and accounting ratios such as earnings per share, rate of return on investment, current ratio, and beta.
3. The amount of energy used per item in a production process.

3. Although this definition is suitable for many business applications, the mathematical theory is more complex and requires a more elaborate definition involving integral calculus (not presented here). We will also refrain from discussing *hybrid* variables, which are neither discrete nor continuous.

Watch Out for Meaningless Numbers

One warning is necessary before you begin analyzing quantitative data: Make sure that the numbers are meaningful! Unfortunately, numbers can be used to record anything at all. If the coding is arbitrary, the results of analysis will be meaningless.

Example

Alphabetical Order of States

Suppose the states of the United States are listed in alphabetical order and coded as 1, 2, 3, …, as follows:

1	Alabama
2	Alaska
3	Arizona
4	Arkansas
⋮	⋮

Now suppose you ask for the average of the states of residence for all employees in your firm's database. The answer is certainly computable. However, the result would be absurd because the numbers assigned to states are not numerically meaningful (although they are convenient for other purposes). To know that the average is 28.35, or somewhere between Nevada and New Hampshire, is not likely to be of use to anybody. The moral is: Be sure that your numbers have meaningful magnitudes before computing with them.

2.3 QUALITATIVE DATA: CATEGORIES

If the data for a variable tells you which one of several nonnumerical categories each item falls into, then the data are **qualitative** (because they record some quality that the item possesses). As you have just seen, care must be taken to avoid the temptation to assign numbers to the categories and then compute with them. If there are just a few categories, then you might work with the percentage of cases in each category (effectively creating something numerical from categorical data). If there are exactly two categories, you can assign the number 0 or 1 to each item and then (for many purposes) continue as if you had quantitative data. But let us first consider the general case in which there are three or more categories.

There are two kinds of qualitative data: *ordinal* (for which there is a meaningful ordering but no meaningful numerical assignment) and *nominal* (for which there is no meaningful order).

Ordinal Qualitative Data

A data set is **ordinal** if there is a meaningful ordering: You can speak of the first (perhaps the "best"), the second, the third, and so on. You can rank the data according to this