Statistics and Variation

CHAPTER 1

Just look at a page from the *Financial Times* website, like the one shown here. It's full of "statistics." Obviously, the writers of the *Financial Times* think all this information is important, but is this what Statistics is all about? Well, yes and no. This page may contain a lot of facts, but as we'll see, the subject is much more interesting and rich than just spreadsheets and tables.

"Why should I learn Statistics?" you might ask. "After all, I don't plan to do this kind of work. In fact, I'm going to hire people to do all of this for me." That's fine. But the decisions you make based on data are too important to delegate. You'll want to be able to interpret the data that surrounds you and to come to your own conclusions. And you'll find that studying Statistics is much more important and enjoyable than you thought.

1.1 So, What Is Statistics?

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

—GEORGE BERNARD SHAW

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

It seems every time we turn around, someone is collecting data on us, from every purchase we make in the grocery store to every click of our mouse as we surf the Web. The United Parcel Service (UPS) tracks every package it ships from one place to another around the world and stores these records in a giant database. You can access part of it if you send or receive a UPS package. The database is about 17 terabytes—about the same size as a database that contained every book in the Library of Congress would be. (But, we suspect, not quite as interesting.) What can anyone hope to do with all these data?

Statistics plays a role in making sense of our complex world. Statisticians assess the risk of genetically engineered foods or of a new drug being considered by the Food and Drug Administration (FDA). Statisticians predict the number of new cases of AIDS by regions of the country or the number of customers likely to respond to a sale at the supermarket. And statisticians help scientists, social scientists, and business leaders understand how unemployment is related to environmental controls, whether enriched early education affects the later performance of school children, and whether vitamin C really prevents illness. Whenever you have data and a need to understand the world, you need Statistics.

If we want to analyze student perceptions of business ethics (a question we'll come back to in a later chapter), should we administer a survey to every single university student in the United States—or, for that matter, in the world? Well, that wouldn't be very practical or cost effective. What should we do instead? Give up and abandon the survey? Maybe we should try to obtain survey responses from a smaller, representative group of students. Statistics can help us make the leap from the data we have at hand to an understanding of the world at large. We talk about the specifics of sampling in Chapter 3, and the theme of generalizing from the specific to the general is one that we revisit throughout this book. We hope this text will empower *you* to draw conclusions from data and make valid business decisions in response to such questions as:

- Do university students from different parts of the world perceive business ethics differently?
- What is the effect of advertising on sales?
- Do aggressive, "high-growth" mutual funds really have higher returns than more conservative funds?
- Is there a seasonal cycle in your firm's revenues and profits?
- What is the relationship between shelf location and cereal sales?
- How reliable are the quarterly forecasts for your firm?
- Are there common characteristics about your customers and why they choose your products?—and, more importantly, are those characteristics the same among those who aren't your customers?

Our ability to answer questions such as these and draw conclusions from data depends largely on our ability to understand *variation*. That may not be the term you expected to find at the end of that sentence, but it is the essence of Statistics. The key to learning from data is understanding the variation that is all around us.

Data vary. People are different. So are economic conditions from month to month. We can't see everything, let alone measure it all. And even what we do measure, we measure imperfectly. So the data we wind up looking at and basing our decisions on provide, at best, an imperfect picture of the world. Variation lies at the heart of what Statistics is all about. How to make sense of it is the central challenge of Statistics.

1.2 How Will This Book Help?

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

Graphs

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

Process

To help you use Statistics to make business decisions, we'll lead you through the entire process of thinking about a problem, finding and showing results, and telling others what you have discovered. The three simple steps to doing Statistics for business right are: **Plan**, **Do**, and **Report**.

PLAN

Plan first. Know where you're headed and why. Clearly defining and understanding your objective will save you a lot of work.

DO

Do is what most students think Statistics is about. The mechanics of calculating statistics and making graphical displays are important, but the computations are usually the least important part of the process. In fact, we usually turn the computations over to technology and get on with understanding what the results tell us.

REPORT

Report what you've learned. Until you've explained your results in a context that someone else can understand, the job is not done.

"Get your facts first, and then you can distort them as much as you please. (Facts are stubborn, but statistics are more pliable.)"

—MARK TWAIN

For Example

At the end of most sections, we present a short example to help you put what you've learned to immediate use. After reading the example, try the corresponding end-of-section exercises at the end of the chapter. These will help prepare you for the other exercises that tend to use all the skills of the chapter.

Guided Example

Each chapter applies the new concepts taught in worked examples called **Guided Examples**. These examples model how you should approach and solve problems using the Plan, Do, Report framework. They illustrate how to plan an analysis, the appropriate techniques to use, and how to report what it all means. These step-by-step examples show you how to produce the kind of solutions and case study reports that instructors and managers or, better yet, clients expect to see. You will find a model solution in the right-hand column and background notes and discussion in the left-hand column.

Just Checking

Sometimes, in the middle of the chapter, you'll find sections called **Just Checking**, which pose a few short questions you can answer without much calculation. Use them to check that you've understood the basic ideas in the chapter. You'll find the answers at the end-of-chapter exercises.

Ethics in Action

Statistics often requires judgment, and the decisions based on statistical analyses may influence people's health and even their lives. Decisions in government can affect policy decisions about how people are treated. In science and industry, interpretations of data can influence consumer safety and the environment. And in business, misunderstanding what the data say can lead to disastrous decisions. The central guiding principle of statistical judgment is the ethical search for a true understanding of the real world. In all spheres of society it is vitally important that a statistical analysis of data be done in an ethical and unbiased way. Allowing preconceived notions, unfair data gathering, or deliberate slanting to affect statistical conclusions is harmful to business and to society.

At various points throughout the book, you will encounter a scenario under the title **Ethics in Action** in which you'll read about an ethical issue. Think about the issue and how you might deal with it. Then read the summary of the issue and one solution to the problem, which follow the scenario. We've related the ethical issues to guidelines that the American Statistical Association has developed.[1] These scenarios can be good topics for discussion. We've presented one solution, but we invite you to think of others.

What Can Go Wrong?

One of the interesting challenges of Statistics is that, unlike some math and science courses, there can be more than one right answer. This is why two statisticians can testify honestly on opposite sides of a court case. And it's why some people think that you can prove anything with statistics. But that's not true. People make mistakes using statistics, and sometimes people misuse statistics to mislead others. Most of the mistakes are avoidable. We're not talking about arithmetic. Mistakes usually involve using a method in the wrong situation or misinterpreting results. So each chapter has a section called **What Can Go Wrong?** to help you avoid some of the most common mistakes that we've seen in our years of consulting and teaching experience.

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

—The Economist, *June 3, 2004*, "*Sloppy stats shame science.*"

Brief CASE

At the end of nearly every chapter you'll find a problem or two that use real data sets and ask you to investigate a question or make a decision. These "brief cases" are a good way to test your ability to attack an open-ended (and thus more realistic) problem. You'll be asked to define the objective, plan your process, complete the analysis, and report your conclusion. These are good opportunities to apply the template provided by the **Guided Examples**. And they provide an opportunity to practice reporting your conclusions in written form to refine your communication skills where statistical results are involved. Data sets for these case studies can be found on the disk included with this text.

[1] http://www.amstat.org/about/ethicalguidelines.cfm

CASE Study

At the end of each section, you'll find a larger project that will help you integrate your knowledge from the entire section you've been studying. These more open-ended projects will help you acquire the skills you'll need to put your knowledge to work in the world of business.

Technology Help: Using the Computer

Although we show you all the formulas you need to understand the calculations, you will most often use a calculator or computer to perform the mechanics of a statistics problem. And the easiest way to calculate statistics with a computer is with a statistics package. Several different statistics packages are used widely. Although they differ in the details of how to use them, they all work from the same basic information and find the same results. Rather than adopt one package for this book, we present generic output and point out common features that you should look for. We also give a table of instructions to get you started on four packages: Excel, Minitab, SPSS, and JMP. Instructions for Excel 2003 and DataDesk can be found on the CD accompanying this textbook.

You'll find all sorts of stuff in margin notes, such as stories and quotations. For example:

"Computers are useless. They can only give you answers."

—Pablo Picasso

While Picasso underestimated the value of good statistics software, he did know that creating a solution requires more than just *Doing*—it means you have to *Plan* and *Report*, too!

From time to time we'll take time out to discuss an interesting or important side issue. We indicate these by setting them apart like this.[2]

What Have We Learned?

At the end of each chapter, you'll see a brief summary of the chapter's learning objectives in a section called **What Have We Learned?** That section also includes a list of the **Terms** you've encountered in the chapter. You won't be able to learn the material from these summaries, but you can use them to check your knowledge of the important ideas in the chapter. If you have the skills, know the terms, and understand the concepts, you should be well prepared—and ready to use Statistics!

Exercises

Beware: No one can learn Statistics just by reading or listening. The only way to learn it is to do it. So, at the end of each chapter (except this one) you'll find **Exercises** designed to help you learn to use the Statistics you've just read about. Some exercises are marked with a red T. You'll find the data for these exercises on the book's website, www.aw-bc.com/sharpe or on the book's disk, so you can use technology as you work the exercises.

We've structured the exercises so that the end-of-section exercises are found first. These can be answered after reading each section. After that you'll find end-of-chapter exercises, designed to help you integrate the topics you've learned in the chapter. We've also paired up and grouped the exercises, so if you're having trouble doing an exercise, you'll find a similar exercise either just before or just after it. You'll find answers to the odd-numbered exercises at the back of the book. But these are only "answers" and not complete solutions. What's the difference? The answers are sketches of the complete solutions. For most problems, your solution

[2] Or in a footnote.

should follow the model of the **Guided Examples**. If your calculations match the numerical parts of the answer and your argument contains the elements shown in the answer, you're on the right track. Your complete solution should explain the context, show your reasoning and calculations, and state your conclusions. Don't worry too much if your numbers don't match the printed answers to every decimal place. Statistics is more than computation—it's about getting the reasoning correct—so pay more attention to how you interpret a result than to what the digit in the third decimal place is.

*Optional Sections and Chapters

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

Getting Started

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

You have to read the book!

Data

CHAPTER 2

Amazon.com

Amazon.com opened for business in July 1995, billing itself even then as "Earth's Biggest Bookstore," with an unusual business plan: They didn't plan to turn a profit for four to five years. Although some shareholders complained when the dotcom bubble burst, Amazon continued its slow, steady growth, becoming profitable for the first time in 2002. Since then, Amazon has remained profitable and has continued to grow. By 2004, they had more than 41 million active customers in over 200 countries and were ranked the 74th most valuable brand by *Business Week*. Their selection of merchandise has expanded to include almost anything you can imagine, from $400,000 necklaces, to yak cheese from Tibet, to the largest book in the world. In 2008, Amazon.com sold nearly $20 billion worth of products online throughout the world.

Amazon R&D is constantly monitoring and evolving their website to best serve their customers and maximize their sales performance. To make changes to the site, they experiment by collecting data and analyzing what works best. As Ronny Kohavi, former director of Data Mining and Personalization, said, "Data trumps intuition. Instead of using our intuition, we experiment on the live site and let our customers tell us what works for them."

Amazon.com has recently stated "many of the important decisions we make at Amazon.com can be made with data. There is a right answer or a wrong answer, a better answer or a worse answer, and math tells us which is which. These are our favorite kinds of decisions."[1] While we might prefer that Amazon refer to these methods as Statistics instead of math, it's clear that data analysis, forecasting, and statistical inference are the core of the decision-making tools of Amazon.com.

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

—RONNY KOHAVI, FORMER DIRECTOR OF DATA MINING AND PERSONALIZATION, AMAZON.COM

Many years ago, stores in small towns knew their customers personally. If you walked into the hobby shop, the owner might tell you about a new bridge that had come in for your Lionel train set. The tailor knew your dad's size, and the hairdresser knew how your mom liked her hair. There are still some stores like that around today, but we're increasingly likely to shop at large stores, by phone, or on the Internet. Even so, when you phone an 800 number to buy new running shoes, customer service representatives may call you by your first name or ask about the socks you bought six weeks ago. Or the company may send an e-mail in October offering new head warmers for winter running. This company has millions of customers, and you called without identifying yourself. How did the sales rep know who you are, where you live, and what you had bought?

The answer to all these questions is data. Collecting data on their customers, transactions, and sales lets companies track inventory and know what their customers prefer. These data can help them predict what their customers may buy in the future so they know how much of each item to stock. The store can use the data and what they learn from the data to improve customer service, mimicking the kind of personal attention a shopper had 50 years ago.

2.1 What *Are* Data?

Businesses have always relied on data for planning and to improve efficiency and quality. Now, more than ever before, businesses rely on the information in data to compete in the global marketplace. Most modern businesses collect information on virtually every transaction performed by the organization, including every item bought or sold. These data are recorded and stored electronically, in vast digital repositories called **data warehouses**.

In the past few decades these data warehouses have grown enormously in size, but with the use of powerful computers, the information contained in them is accessible and used to help make decisions, sometimes almost instantaneously. When you pay with your credit card, for example, the information about the transaction is transmitted to a central computer where it is processed and analyzed. A decision whether to approve or deny your purchase is made and transmitted back to the point of sale, all within a few seconds.

[1]Amazon.com 2008 Annual Report

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

Leading companies are embracing business analytics. Richard Fairbank, the CEO and founder of Capital One, revolutionized the credit card industry by realizing that credit card transactions hold the key to understanding customer behavior. Reed Hastings, a former computer science major, is the founder and CEO of Netflix. Netflix uses analytics on customer information both to recommend new movies and to adapt the website that customers see to individual tastes. Netflix offered a $1 million prize to anyone who could improve on the accuracy of their recommendations by more than 10%. That prize was won in 2009 by a team of statisticians and computer scientists using predictive analytics and data-mining techniques. The Oakland Athletics use analytics to judge players instead of the traditional methods used by scouts and baseball experts for over a hundred years. The book *Moneyball* documents how business analytics enabled them to put together a team that could compete against the richer teams in spite of the severely limited resources available to the front office.

To understand better what data are, let's look at some hypothetical company records that Amazon might collect:

105-2686834-3759466	B0000010AA	10.99	Chris G.	902	Boston	15.98	Kansas	Illinois
Samuel P.	Orange County	105-9318443-4200264	105-1872500-0198646	N	B000068ZVQ	Bad Blood	Nashville	Katherine H.
Canada	Garbage	16.99	Ohio	N	Chicago	N	11.99	Massachusetts
B000002BK9	312	Monique D.	Y	413	B00000I5Y6	440	103-2628345-9238664	Let Go

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

THE W'S:
WHO
WHAT
WHEN
WHERE
WHY

Try to guess what these data represent. Why is that hard? Because these data have no *context*. Whether the data are numerical (consisting only of numbers), alphabetic (consisting only of letters), or alphanumerical (mixed numbers and letters), they are useless unless we know what they represent. Newspaper journalists know that the lead paragraph of a good story should establish the "Five W's": *who*, *what*, *when*, *where*, and (if possible) *why*. Often, we add *how* to the list as well. Answering these questions can provide a **context** for data values and make them meaningful. The answers to the first two questions are essential. If you can't answer *who* and *what*, you don't have data, and you don't have any useful information.

We can make the meaning clear if we add the context of ***who*** the data are about and *what* was measured and organize the values into a **data table** such as this one.

Order Number	Name	State/Country	Price	Area Code	Previous Album Download	Gift?	ASIN	Artist
105-2686834-3759466	Katherine H.	Ohio	10.99	440	Nashville	N	B00000I5Y6	Kansas
105-9318443-4200264	Samuel P.	Illinois	16.99	312	Orange County	Y	B000002BK9	Boston
105-1872500-0198646	Chris G.	Massachusetts	15.98	413	Bad Blood	N	B000068ZVQ	Chicago
103-2628345-9238664	Monique D.	Canada	11.99	902	Let Go	N	B0000010AA	Garbage
002-1663369-6638649	Katherine H.	Ohio	10.99	440	Best of Kansas	N	B002MXA7Q0	Kansas

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

Look at the rows of Table 2.2. Now we can see that these are five purchase records, relating to album downloads from Amazon. In general, the rows of a data table correspond to individual **cases** about which we've recorded some characteristics called **variables**.

Cases go by different names, depending on the situation. Individuals who answer a survey are referred to as **respondents**. People on whom we experiment are **subjects** or (in an attempt to acknowledge the importance of their role in the experiment) **participants**, but animals, plants, websites, and other inanimate subjects are often called **experimental units**. Often we call cases just what they are: for example, *customers*, *economic quarters*, or *companies*. In a database, rows are called **records**—in this example, purchase records. Perhaps the most generic term is **cases**. In Table 2.2, the cases are the individual orders.

The column titles (variable names) tell *what* has been recorded. What does a row of Table 2.2 represent? Be careful. Even if people are involved, the cases may not correspond to people. For example, in Table 2.2, each row is a different order and not the customer who made the purchases (notice that the same person made two different orders). A common place to find the *who* of the table is the leftmost column. It's often an identifying variable for the cases, in this example, the order number.

If you collect the data yourself, you'll know what the cases are and how the variables are defined. But, often, you'll be looking at data that someone else collected. The information about the data, called the metadata, might have to come from the company's database administrator or from the *information technology* department of a company. **Metadata** typically contains information about *how*, *when*, and *where* (and possibly *why*) the data were collected; *who* each case represents; and the definitions of all the variables.

A general term for a data table like the one shown in Table 2.2 is a **spreadsheet**, a name that comes from bookkeeping ledgers of financial information. The data were typically spread across facing pages of a bound ledger, the book used by an accountant for keeping records of expenditures and sources of income. For the accountant, the columns were the types of expenses and income, and the cases were transactions, typically invoices or receipts. These days, it is common to keep modest-size datasets in a spreadsheet even if no accounting is involved. It is usually easy to move a data table from a spreadsheet program to a program designed for statistical graphics and analysis, either directly or by copying the data table and pasting it into the statistics program.

Although data tables and spreadsheets are great for relatively small data sets, they are cumbersome for the complex data sets that companies must maintain on a day-to-day basis. Try to imagine a spreadsheet from Amazon with customers in the rows and products in the columns. Amazon has tens of millions of customers and hundreds of thousands of products. But very few customers have purchased more than a few dozen items, so almost all the entries would be blank—not a very

efficient way to store information. For that reason, various other database architectures are used to store data. The most common is a relational database.

In a **relational database**, two or more separate data tables are linked together so that information can be merged across them. Each data table is a *relation* because it is about a specific set of cases with information about each of these cases for all (or at least most) of the variables ("fields" in database terminology). For example, a table of customers, along with demographic information on each, is such a relation. A data table with information about a different collection of cases is a different relation. For example, a data table of all the items sold by the company, including information on price, inventory, and past history, is a relation as well (for example, as in Table 2.3). Finally, the day-to-day transactions may be held in a third database where each purchase of an item by a customer is listed as a case. In a relational database, these three relations can be linked together. For example, you can look up a customer to see what he or she purchased or look up an item to see which customers purchased it.

In statistics, all analyses are performed on a single data table. But often the data must be retrieved from a relational database. Retrieving data from these databases often requires specific expertise with that software. In the rest of the book, we'll assume that all data have been downloaded to a data table or spreadsheet with variables listed as columns and cases as the rows.

Customers

Customer Number	Name	City	State	Zip Code	Customer since	Gold Member?
473859	R. De Veaux	Williamstown	MA	01267	2007	No
127389	N. Sharpe	Washington	DC	20052	2000	Yes
335682	P. Velleman	Ithaca	NY	14580	2003	No
...						

Items

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

Transactions

Transaction Number	Date	Customer Number	Product ID	Quantity	Shipping Method	Free Ship?
T23478923	9/15/08	473859	SC5662	1	UPS 2nd Day	N
T23478924	9/15/08	473859	TH2839	1	UPS 2nd Day	N
T63928934	10/20/08	335682	TH2839	3	UPS Ground	N
T72348299	12/22/08	127389	RS3883	1	Fed Ex Ovnt	Y

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

For Example — Identifying variables and the W's

Carly, a marketing manager at a credit card bank, wants to know if an offer mailed 3 months ago has affected customers' use of their cards. To answer that, she asks the information technology department to assemble the following information for each customer: total spending on the card during the 3 months before the offer (*Pre Spending*); spending for 3 months after the offer (*Post Spending*); the customer's *Age* (by category); what kind of expenditure they made (*Segment*); if customers are enrolled in the website (*Enroll?*); what offer they were sent (*Offer*); and the amount each customer has spent on the card in their segment (*Segment Spend*). She gets a spreadsheet whose first six rows look like this:

Account ID	Pre Spending	Post Spending	Age	Segment	Enroll?	Offer	Segment Spend
393371	$2,698.12	$6,261.40	25–34	Travel/Ent	NO	None	$887.36
462715	$2,707.92	$3,397.22	45–54	Retail	NO	Gift Card	$5,062.55
433469	$800.51	$4,196.77	65+	Retail	NO	None	$673.80
462716	$3,459.52	$3,335.00	25–34	Services	YES	Double Miles	$800.75
420605	$2,106.48	$5,576.83	35–44	Leisure	YES	Double Miles	$3,064.81
473703	$2,603.92	$7,397.50	<25	Travel/Ent	YES	Double Miles	$491.29

Question: Identify the cases and the variables. Describe as many of the W's as you can for this data set.

Answer: The cases are individual customers of the credit card bank. The data are from the internal records of the credit card bank from the past 6 months (3 months before and 3 months after an offer was sent to the customers). The variables include the account ID of the customer (*Account ID*) and the amount charged on the card before (*Pre Spending*) and after (*Post Spending*) the offer was sent out. Also included are the customer's *Age*, marketing *Segment*, whether they enrolled on the website (*Enroll?*), what offer they were sent (*Offer*), and how much they charged on the card in their marketing segment (*Segment Spend*).

2.2 Variable Types

Variables play different roles, and knowing the variable's *type* is crucial to knowing what to do with it and what it can tell us. When a variable names categories and answers questions about how cases fall into those categories, we call it a **categorical**, or **qualitative, variable**. When a variable has measured numerical values with *units* and the variable tells us about the quantity of what is measured, we call it a **quantitative variable**. Classifying a variable into categorical or quantitative can help us decide what to do with a variable, but doing so is often more about what we hope to learn from a variable than about the variable itself. It's the questions we ask of a variable (the *why* of our analysis) that shape how we think about it and how we treat it.

Descriptive responses to questions are often categories. For example, the responses to the questions "What type of mutual fund do you invest in?" or "What kind of advertising does your firm use?" yield categorical values. An important special case of categorical variables is one that has only two possible responses (usually "yes" or "no"), which arise naturally from questions like "Do you invest in the stock market?" or "Do you make online purchases from this website?"

If the variable has values that are not numbers, it's clearly categorical (or needs to be recoded). However, if the values are numbers, you need to be careful. It may be considered quantitative if the values actually measure a quantity of something. Otherwise, it's categorical. For example, area codes are numbers, but the numerical values of area codes don't have numerical meaning (see the side bar). The numbers assigned by the area codes are codes that *categorize* the phone number into a geographical area. So, we treat area code as a categorical variable.

For quantitative variables, the **units** tell how each value has been measured. Even more important, units such as yen, cubits, carats, angstroms, nanoseconds, miles per hour, or degrees Celsius tell us the *scale* of measurement, so we know how far apart two values are. Without units, the values of a measured variable have no meaning. It does little good to be promised a raise of 5000 a year if you don't know whether it will be paid in euros, dollars, yen, or Estonian krooni. An essential part of a quantitative variable is its units.

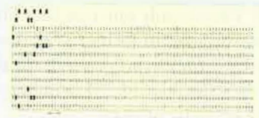
Categorical, or Quantitative?

It is wise to be careful. The *what* and *why* of area codes are not as simple as they may first seem.

When area codes were first introduced all phones had dials. To reduce wear and tear on the dials and to speed the most number of calls, the lowest-digit codes (the easiest to dial) were assigned to the largest cities. So, New York City was given 212, Chicago 312, LA 213 and Philadelphia 215, but rural upstate New York was 607, Joliet was 815, and San Diego 619. Back then, the numerical value of an area code could be used to guess something about the population of its region. But after dials gave way to push buttons, new area codes were assigned without regard to population and area codes are now just categories.

Variable Names that Make Sense

One tradition that still hangs on in some places is to name variables with cryptic abbreviations in uppercase letters. This can be traced back to the 1960s, when computer programs were controlled with instructions punched on cards. The earliest punch card equipment used only uppercase letters, and statistics programs limited variable names to six or eight characters, so variables had names like PRSRF3. Modern programs don't have such restrictive limits, so there is no reason not to use names that make sense.



Sometimes the type of the variable is clear. But some variables can answer both kinds of questions and how they are classified depends on their use. For example, the variable *Age* would be considered quantitative if the responses were numerical and they had units. A doctor would certainly need *Age* to be quantitative. The units could be years, but for infants, the doctor would want even more precise units, like months, or even days. On the other hand, if Amazon asked your *Age*, it might lump together the values into categories like "Child (12 years or less)," "Teen (13 to 19)," "Adult (20 to 64)," or "Senior (65 or over)." For many purposes, like knowing which CD ad to send you, that might be all the information Amazon might need. In this case, Amazon has made *Age* a categorical variable.

Question	Categories or Responses
Do you invest in the stock market?	__ Yes __ No
What kind of advertising do you use?	__ Newspapers __ Internet __ Direct mailings
What is your class at school?	__ Freshman __ Sophomore __ Junior __ Senior
I would recommend this course to another student.	__ Strongly Disagree __ Slightly Disagree __ Slightly Agree __ Strongly Agree
How satisfied are you with this product?	__ Very Unsatisfied __ Unsatisfied __ Satisfied __ Very Satisfied

Table 2.4 Some examples of categorical variables.

When Amazon considers a special offer of free shipping to customers, they might first analyze how purchases have been shipped in the recent past. They might start by counting the number of purchases shipped in each category: ground transportation, second-day air, and overnight air. Counting is a natural way to summarize a categorical variable like *Shipping Method*. Chapter 4 discusses summaries and displays of categorical variables more fully. Chapter 5 discusses quantitative variables, which require different summaries and displays.

Shipping Method	Number of Purchases
Ground	20,345
Second-day	7890
Overnight	5432

Table 2.5 A summary of the categorical variable *Shipping Method* that shows the counts, or number of cases for each category.

Identifiers

What's your student ID number? It may be numerical, but is it a quantitative variable? No, it doesn't have units. Is it categorical? Yes, but a special kind. Look at how many categories there are and at how many individuals there are in each category. There are exactly as many categories as individuals and only one individual in each category. While it's easy to count the totals for each category, it's not very interesting. This is an **identifier variable**. Amazon wants to know who you are when you sign in again and doesn't want to confuse you with some other customer. So they assign you a unique identifier.

Identifier variables themselves don't tell us anything useful about the categories because we know there is exactly one individual in each. However, they are crucial in this era of large data sets because by uniquely identifying the cases, they make it possible to combine data from different sources, protect confidentiality, and provide unique labels. Most company databases are, in fact, relational databases. The identifier is crucial to linking one data table to another in a relational database. The identifiers in Table 2.3 are the *Customer Number*, *Product ID*, and *Transaction Number*. Variables like *UPS Tracking Number*, *Social Security Number*, and Amazon's *ASIN* are other examples of identifiers.

You'll want to recognize when a variable is playing the role of an identifier so you won't be tempted to analyze it. Knowing that Amazon's average ASIN number increased 10% from 2007 to 2008 doesn't really tell you anything—any more than analyzing any categorical variable as if it were quantitative would.

Be careful not to be inflexible in your typing of variables. Variables can play different roles, depending on the question we ask of them, and classifying variables rigidly into types can be misleading. For example, in their annual reports, Amazon refers to its database and looks at the variables *Sales* and *Year*. When analysts ask how many books Amazon sold in 2005, what role does *Year* play? There's only one row for 2005, and *Year* identifies it, so it plays the role of an identifier variable. In its role as an identifier, you might match other data from Amazon, or the economy in general, for the same year. But analysts also track sales growth over time. In this role, *Year* measures time. Now it's being treated as a quantitative variable with units of years.

Other Data Types

A survey might ask:

"How satisfied were you with the service you received?"

1) Not satisfied; 2) Somewhat satisfied; 3) Moderately satisfied; or 4) Extremely satisfied.

Is this variable categorical or quantitative? There is certainly an *order* of perceived worth; higher numbers indicate higher perceived worth. An employee whose customer responses average around 4 seems to be doing a better job than one whose averages are around 2, but are they *twice* as good? Because the values are not strictly numbers, we can't really say and so we should be careful about treating *Customer Satisfaction* as purely quantitative. When, as in this example, the values of a categorical value have an intrinsic order, we can say that the categorical variable is **ordinal**. By contrast, a categorical variable that names categories that don't have order is sometimes called **nominal**. Values can be individually ordered (e.g., the ranks of employees based on the number of days they've worked for the company) or ordered in classes (e.g., Freshman, Sophomore, Junior, Senior). Ordering is not absolute; how the values are ordered depends on the purpose of the ordering. For example, are the categories Infant, Youth, Teen, Adult, and Senior ordinal? Well, if we are ordering on age, they surely are and how to order the categories is clear. But if we are ordering (as Amazon might) on purchase volume, it is likely that either Teen or Adult will be the top group.[2]

Cross-Sectional and Time Series Data

The quantitative variable *Total Revenue* in Table 2.6 is an example of a time series. A **time series** is a single variable measured at regular intervals over time. Time series are common in business. Typical measuring points are months, quarters, or years, but virtually any consistently-spaced time interval is possible. Variables collected over time hold special challenges for statistical analysis, and Chapter 20 discusses these in more detail.

[2]Some people differentiate quantitative variables according to whether their measured values have a defined value for zero. This is a technical distinction and usually not one we'll need to make. (For example, it isn't correct to say that a temperature of 80°F is twice as hot as 40°F because 0° is an arbitrary value. On the Celsius scale those temperatures are 26.67°C and 4.44°C—a ratio of 6.) The term *interval scale* is sometimes applied to data such as these, and the term *ratio scale* is applied to measurements for which such ratios are appropriate.

Year	Total Revenue (in $M)
2002	3288.9
2003	4075.5
2004	5294.2
2005	6369.3
2006	7786.9
2007	9441.5
2008	10,383.0
2009	9774.6

Table 2.6 Starbucks's total revenue (in $M) for the years 2002 to 2009.

By contrast, most of the methods in this book are better suited for **cross-sectional data**, where several variables are measured at the same time point. On the other hand, if we collect data on sales revenue, number of customers, and expenses for last month at *each* Starbucks (more than 16,000 locations as of 2010) at one point in time, this would be cross-sectional data. Cross-sectional data may contain some time information (such as dates), but it isn't a time series because it isn't measured at regular intervals. Because different methods are used to analyze these different types of data, it is important to be able to identify both time series and cross-sectional data sets.

For Example — Identifying the types of variables

Question: Before she can continue with her analysis, Carly (from the example on page 12) must classify each variable as being quantitative or categorical (or possibly both), and whether the data are a time series or cross-sectional. For quantitative variables, what are the units? For categorical variables, are they nominal or ordinal?

Answer:

Account ID – categorical (nominal, identifier)

Pre Spending – quantitative (units $)

Post Spending – quantitative (units $)

Age – categorical (ordinal). Could be quantitative if we had more precise information

Segment – categorical (nominal)

Enroll? – categorical (nominal)

Offer – categorical (nominal)

Segment Spend – quantitative (units $)

The data are cross-sectional. We do not have successive values of a single variable over time.

2.3 Data Sources: Where, How, and When

We must know *who*, *what*, and *why* to analyze data. Without knowing these three, we don't have enough to start. Of course, we'd always like to know more because the more we know, the more we'll understand. If possible, we'd like to know the *where*, *how*, and *when* of data as well. Values recorded in 1947 may mean something different than similar values recorded last year. Values measured in Abu Dhabi may differ in meaning from similar measurements made in Mexico.

How the data are collected can make the difference between insight and nonsense. As we'll see later, data that come from a voluntary survey on the Internet are almost always worthless. In a recent Internet poll, 84% of respondents said "no" to the question of whether subprime borrowers should be bailed out. While it may be true that 84% of those 23,418 respondents did say that, it's dangerous to assume that that group is representative of any larger group. To make inferences from the data you have at hand to the world at large, you need to ensure that the data you have are representative of the larger group. Chapter 3 discusses sound methods for *designing* a *survey* or poll to help ensure that the inferences you make are valid.

Another way to collect valid data is by performing an experiment in which you actively manipulate variables (called factors) to see what happens. Most of the "junk mail" credit card offers that you receive are actually experiments done by marketing groups in those companies. They may make different versions of an offer to selected groups of customers to see which one works best before rolling out the winning idea to the entire customer base. Chapter 22 discusses both the design and the analysis of experiments like these.

Sometimes, the answer to the question you have may be found in data that someone, or more typically, some organization has already collected. Internally, companies may analyze data from their own data bases or data warehouse. They may also supplement or rely entirely on data collected by others. Many companies, nonprofit organizations, and government agencies collect vast amounts of data via the Internet. Some organizations may charge a fee for accessing or downloading their data. The U.S. government collects information on nearly every aspect of life in the United States, both social and economic (see for example www.census.gov, or more generally, www.usa.gov), as the European Union does for Europe (see ec.europa.eu/eurostat). International organizations such as the World Health Organization (www.who.org) and polling agencies such as Gallup (www.gallup.com) offer information on a variety of topics as well. Data like these typically do not come from a designed survey or experiment. They are most often collected for different purposes than the analysis you may want to perform. Although they are plentiful, you should be careful when generalizing from data like these. Information about how, when, where, and why the data were collected may not be available. Unless the data were collected in a way that ensures that they are representative of the population in which you are interested, you may be misled if you try to draw conclusions from them. Chapter 24 discusses data mining, which attempts to use large amounts of "found" data to make hypotheses and draw insights. While it can be tempting, interesting, and even useful to analyze such happenstance data, remember that the only way to be sure that a generalization is valid is if the data come from a properly designed survey or experiment.

There's a world of data on the Internet

These days, one of the richest sources of data is the Internet. With a bit of practice, you can learn to find data on almost any subject. We found many of the data sets used in this book by searching on the Internet. The Internet has both advantages and disadvantages as a source of data. Among the advantages are the fact that often you'll be able to find even more current data than we present. One disadvantage is that references to Internet addresses can "break" as sites evolve, move, and die. Another disadvantage is that important metadata—information about the collection, quality, and intent of the data—may be missing.

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

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

Throughout this book, whenever we introduce data, we will provide a margin note listing some of the W's of the data and, where possible, offer a reference for the source of the data. It's a habit we recommend. The first step of any data analysis is to know why you are examining the data (what you want to know), whom each row of your data table refers to, and what the variables (the columns of the table) record. These are the *Why*, the *Who*, and the *What*. Identifying them is a key part of the *Plan* step of any analysis. Make sure you know all three before you spend time analyzing the data.

For Example Identifying data sources

On the basis of her initial analysis, Carly asks her colleague Ying Mei to e-mail a sample of customers from the Travel and Entertainment segment and ask about their card use and household demographics. Carly asks another colleague, Gregg, to design a study about their double miles offer. In this study, a random sample of customers receives one of three offers: the standard double miles offer; a double miles offer good on any airline; or no offer.

Question: For each of the three data sets—Carly's original data set and Ying Mei's and Gregg's sets—state whether they come from a designed survey or a designed experiment or are collected in another way.

Answer: Carly's data set was derived from transactional data, not part of a survey or experiment. Ying Mei's data come from a designed survey, and Gregg's data come from a designed experiment.

Just Checking

An insurance company that specializes in commercial property insurance has a separate database for their policies that involve churches and schools. Here is a small portion of that database.

Policy Number	Years Claim Free	Net Property Premium ($)	Net Liability Premium ($)	Total Property Value ($1,000)	Median Age in Zip Code	School?	Territory	Coverage
4000174699	1	3107	503	1036	40	FALSE	AL580	BLANKET
8000571997	2	1036	261	748	42	FALSE	PA192	SPECIFIC
8000623296	1	438	353	344	30	FALSE	ID60	BLANKET
3000495296	1	582	339	270	35	TRUE	NC340	BLANKET
5000291199	4	993	357	218	43	FALSE	OK590	BLANKET
8000470297	2	433	622	108	31	FALSE	NV140	BLANKET
1000042399	4	2461	1016	1544	41	TRUE	NJ20	BLANKET
4000554596	0	7340	1782	5121	44	FALSE	FL530	BLANKET
3000260397	0	1458	261	1037	42	FALSE	NC560	BLANKET
8000333297	2	392	351	177	40	FALSE	OR190	BLANKET
4000174699	1	3107	503	1036	40	FALSE	AL580	BLANKET

1 List as many of the W's as you can for this data set.

2 Classify each variable as to whether you think it should be treated as categorical or quantitative (or both); if quantitative, identify the units.

What Can Go Wrong?

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

Ethics in Action

Sarah Potterman, a doctoral student in educational psychology, is researching the effectiveness of various interventions recommended to help children with learning disabilities improve their reading skills. Among the approaches examined is an interactive software system that uses analogy-based phonics. Sarah contacted the company that developed this software, RSPT Inc., in order to obtain the system free of charge for use in her research. RSPT Inc. expressed interest in having her compare their product with other intervention strategies and was quite confident that their approach would be the most effective. Not only did the company provide Sarah with free software, but RSPT Inc. also generously offered to fund her research with a grant to cover her data collection and analysis costs.

ETHICAL ISSUE *Both the researcher and company should be careful about the funding source having a vested interest in the research result (related to Item H, ASA Ethical Guidelines).*

ETHICAL SOLUTION *RSPT Inc. should not pressure Sarah Potterman to obtain a particular result. Both parties should agree on paper before the research is begun that the research results can be published even if they show that RSPT's interactive software system is not the most effective.*

Jim Hopler is operations manager for a local office of a top-ranked full-service brokerage firm. With increasing competition from both discount and online brokers, Jim's firm has redirected attention to attaining exceptional customer service through its client-facing staff, namely brokers. In particular, they wish to emphasize the excellent advisory services provided by their brokers. Results from surveying clients about the advice received from brokers at the local office revealed that 20% rated it *poor*, 5% rated it *below average*, 15% rated it *average*, 10% rated it *above average,* and 50% rated it *outstanding*. With corporate approval, Jim and his management team instituted several changes in an effort to provide the best possible advisory services at the local office. Their goal was to increase the percentage of clients who viewed their advisory services as *outstanding.* Surveys conducted after the changes were implemented showed the following results: 5% *poor*, 5% *below average*, 20% *average*, 40% *above average,* and 30% *outstanding*. In discussing these results, the management team expressed concern that the percentage of clients who considered their advisory services *outstanding* fell from 50% to 30%. One member of the team suggested an alternative way of summarizing the data. By coding the categories on a scale from 1 = poor to 5 = outstanding and computing the average, they found that the average rating increased from 3.65 to 3.85 as a result of the changes implemented. Jim was delighted to see that their changes were successful in improving the level of advisory services offered at the local office. In his report to corporate, he only included average ratings for the client surveys.

ETHICAL ISSUE *By taking an average, Jim is able to show improved customer satisfaction. However, their goal was to increase the percentage of outstanding ratings. Jim redefined his study after the fact to support a position (related to Item A, ASA Ethical Guidelines).*

ETHICAL SOLUTION *Jim should report the percentages for each rating category. He can also report the average. He may wish to include in his report a discussion of what those different ways of looking at the data say and why they appear to differ. He may also want to explore with the survey participants the perceived differences between "above average" and "outstanding."*

What Have We Learned?

Learning Objectives

- Understand that data are values, whether numerical or labels, together with their context.
  - *who*, *what*, *why*, *where*, *when* (and *how*)—the W's—help nail down the context of the data.
  - We must know *who*, *what*, and *why* to be able to say anything useful based on the data. The *who* are the cases. The *what* are the variables. A variable gives information about each of the cases. The *why* helps us decide which way to treat the variables.
  - Stop and identify the W's whenever you have data, and be sure you can identify the cases and the variables.
- Identify whether a variable is being used as categorical or quantitative.
  - Categorical variables identify a category for each case. Usually we think about the counts of cases that fall in each category. (An exception is an identifier variable that just names each case.)
  - Quantitative variables record measurements or amounts of something; they must have units.
  - Sometimes we may treat the same variable as categorical or quantitative depending on what we want to learn from it, which means some variables can't be pigeonholed as one type or the other.
- Consider the source of your data and the reasons the data were collected. That can help you understand what you might be able to learn from the data.

Terms

Business analytics The process of using statistical analysis and modeling to drive business decisions.

Case A case is an individual about whom or which we have data.

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

Context The context ideally tells *who* was measured, *what* was measured, *how* the data were collected, *where* the data were collected, and *when* and *why* the study was performed.

Cross-sectional data Data taken from situations that vary over time but measured at a single time instant is said to be a cross-section of the time series.

Data Recorded values whether numbers or labels, together with their context.

Data mining The process of using a variety of statistical tools to analyze large data bases or data warehouses.

Data table An arrangement of data in which each row represents a case and each column represents a variable.

Data warehouse A large data base of information collected by a company or other organization usually to record transactions that the organization makes, but also used for analysis via data mining.

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

Identifier variable A categorical variable that records a unique value for each case, used to name or identify it.

Metadata Auxiliary information about variables in a database, typically including *how*, *when*, and *where* (and possibly *why*) the data were collected; *who* each case represents; and the definitions of all the variables.

Nominal variable The term "nominal" can be applied to a variable whose values are used only to name categories.

Ordinal variable The term "ordinal" can be applied to a variable whose categorical values possess some kind of order.

Participant	A human experimental unit. Also called a subject.
Quantitative variable	A variable in which the numbers are values of measured quantities with units.
Record	Information about an individual in a database.
Relational database	A relational database stores and retrieves information. Within the database, information is kept in data tables that can be "related" to each other.
Respondent	Someone who answers, or responds to, a survey.
Spreadsheet	A spreadsheet is layout designed for accounting that is often used to store and manage data tables. Excel is a common example of a spreadsheet program.
Subject	A human experimental unit. Also called a participant.
Time series	Data measured over time. Usually the time intervals are equally spaced or regularly spaced (e.g., every week, every quarter, or every year).
Transactional Data	Data collected to record the individual transactions of a company or organization.
Units	A quantity or amount adopted as a standard of measurement, such as dollars, hours, or grams.
Variable	A variable holds information about the same characteristic for many cases.

Technology Help: Data on the Computer

Most often we find statistics on a computer using a program, or *package*, designed for that purpose. There are many different statistics packages, but they all do essentially the same things. If you understand what the computer needs to know to do what you want and what it needs to show you in return, you can figure out the specific details of most packages pretty easily.

For example, to get your data into a computer statistics package, you need to tell the computer:

- Where to find the data. This usually means directing the computer to a file stored on your computer's disk or to data on a database. Or it might just mean that you have copied the data from a spreadsheet program or Internet site and it is currently on your computer's clipboard. Usually, the data should be in the form of a data table. Most computer statistics packages prefer the *delimiter* that marks the division between elements of a data table to be a *tab* character and the delimiter that marks the end of a case to be a *return* character.
- Where to put the data. (Usually this is handled automatically.)
- What to call the variables. Some data tables have variable names as the first row of the data, and often statistics packages can take the variable names from the first row automatically.

Brief CASE

Credit Card Bank

Like all credit and charge card companies, this company makes money on each of its cardholders' transactions. Thus, its profitability is directly linked to card usage. To increase customer spending on its cards, the company sends many different offers to its cardholders, and market researchers analyze the results to see which offers yield the largest increases in the average amount charged.

On your disk (in the file **Credit_Card_Bank**) is part of a database like the one used by the researchers. For each customer, it contains several variables in a spreadsheet.

Examine the data in the data file. List as many of the W's as you can for these data and classify each variable as categorical or quantitative. If quantitative, identify the units.

Exercises

SECTION 2.1

1. A real estate major collected information on some recent local home sales. The first 6 lines of the database appear below. The columns correspond to the house identification number, the community name, the zip code, the number of acres of the property, the year the house was built, the market value, and the size of the living area (in square feet).

a) What does a row correspond to in this data table? How would you best describe its role: as a participant, subject, case, respondent, or experimental unit?
b) How many variables are measured on each row?

HOUSE_ID	NEIGHBORHOOD	MAIL_ZIP	ACRES	YR_BUILT	FULL_MARKET_VALUE	SFLA
413400536	Greenfield Manor	12859	1.00	1967	100400	960
4128001474	Fort Amherst	12801	0.09	1961	132500	906
412800344	Dublin	12309	1.65	1993	140000	1620
4128001552	Granite Springs	10598	0.33	1969	67100	900
412800352	Arcady	10562	2.29	1955	190000	1224
413400322	Ormsbee	12859	9.13	1997	126900	1056

2. A local bookstore is keeping a database of its customers to find out more about their spending habits and so that the store can start to make personal recommendations based on past purchases. Here are the first five rows of their database:

a) What does a row correspond to in this data table? How would you best describe its role: as a participant, subject, case, respondent, or experimental unit?
b) How many variables are measured on each row?

Transaction ID	Customer ID	Date	ISBN Number of Purchase	Price	Coupon?	Gift?	Quantity
29784320912	4J438	11/12/2009	345-23-2355	$29.95	N	N	1
26483589001	3K729	9/30/2009	983-83-2739	$16.99	N	N	1
26483589002	3K729	9/30/2009	102-65-2332	$9.95	Y	N	1
36429489305	3U034	12/5/2009	295-39-5884	$35.00	N	Y	1
36429489306	3U034	12/5/2009	183-38-2957	$79.95	N	Y	1

SECTION 2.2

3. Referring to the real estate data table of Exercise 1,

a) For each variable, would you describe it as primarily categorical, or quantitative? If quantitative, what are the units? If categorical, is it ordinal or simply nominal?
b) Are these data a time series, or are these cross-sectional? Explain briefly.

4. Referring to the bookstore data table of Exercise 2,

a) For each variable, would you describe it as primarily categorical, or quantitative? If quantitative, what are the units? If categorical, is it ordinal or simply nominal?
b) Are these data a time series, or are these cross-sectional? Explain briefly.

SECTION 2.3

5. For the real estate data of Exercise 1, do the data appear to have come from a designed survey or experiment? What concerns might you have about drawing conclusions from this data set?

6. A student finds data on an Internet site that contains financial information about selected companies. He plans to analyze the data and use the results to develop a stock investment strategy. What kind of data source is he using? What concerns might you have about drawing conclusions from this data set?

CHAPTER EXERCISES

For each description of data in Exercises 7 to 26, identify the W's, name the variables, specify for each variable whether its use indicates it should be treated as categorical or quantitative, and for any quantitative variable identify the units in which it was measured (or note that they were not provided). Specify whether the data come from a designed survey or experiment. Are the variables time series or cross-sectional? Report any concerns you have as well.

closed (in $), supervisor's rating (1–10), years with the company.

29. Company performance. Data collected for financial planning: weekly sales, week (week number of the year), sales predicted by last year's plan, difference between predicted sales and realized sales.

30. Command performance. Data collected on investments in Broadway shows: number of investors, total invested, name of the show, profit/loss after one year.

For the following examples in Exercises 31 to 34, indicate whether the data are a time series or a cross section.

31. Car sales. Number of cars sold by each salesperson in a dealership in September.

32. Motorcycle sales. Number of motorcycles sold by a dealership in each month of 2008.

33. Cross sections. Average diameter of trees brought to a sawmill in each week of a year.

34. Series. Attendance at the third World Series game recording the age of each fan.

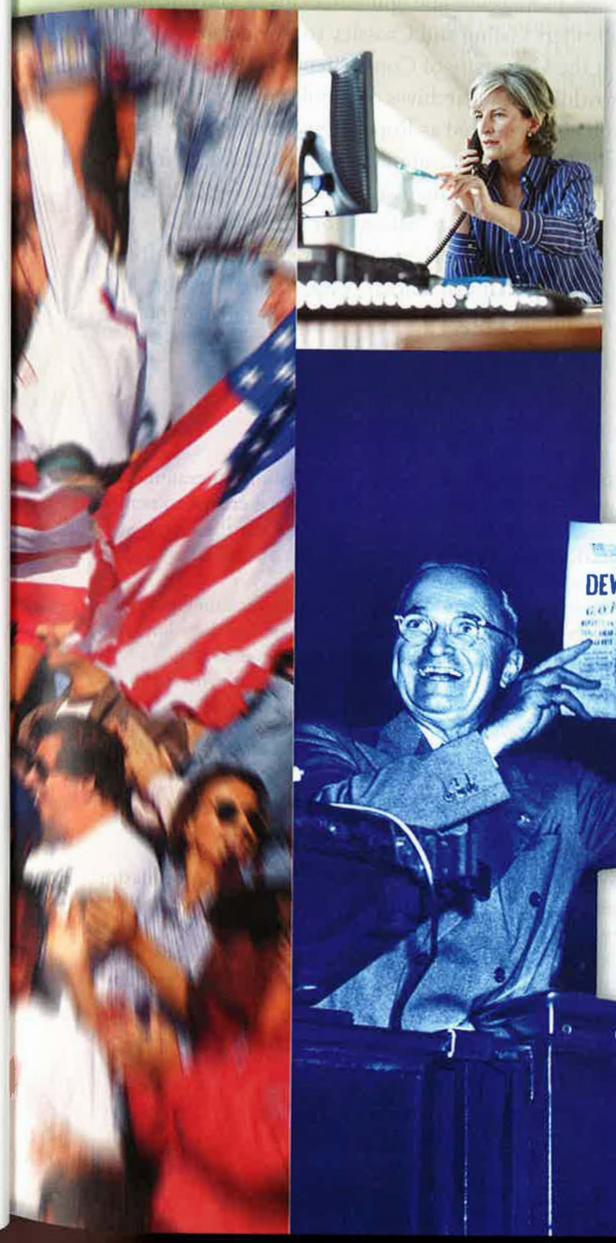
Just Checking Answers

1 Who—policies on churches and schools
What—policy number, years claim free, net property premium ($), net liability premium ($), total property value ($000), median age in zip code, school?, territory, coverage
How—company records
When—not given

2 Policy number: identifier (categorical)
Years claim free: quantitative
Net property premium: quantitative ($)
Net liability premium: quantitative ($)
Total property value: quantitative ($)
Median age in zip code: quantitative
School?: categorical (true/false)
Territory: categorical
Coverage: categorical

Surveys and Sampling

CHAPTER 3



Roper Polls

Public opinion polls are a relatively new phenomenon. In 1948, as a result of telephone surveys of likely voters, all of the major organizations—Gallup, Roper, and Crossley—consistently predicted, throughout the summer and into the fall, that Thomas Dewey would defeat Harry Truman in the November presidential election. By October the results seemed so clear that *Fortune* magazine declared, "Due to the overwhelming evidence, *Fortune* and Mr. Roper plan no further detailed reports on change of opinion in the forthcoming presidential campaign. . . ."

Of course, Harry Truman went on to win the 1948 election, and the picture of Truman in the early morning after the election holding up the *Chicago Tribune* (printed the night before), with its headline declaring Dewey the winner, has become legend.

The public's faith in opinion polls plummeted after the election, but Elmo Roper vigorously defended the pollsters. Roper was a principal and founder of one of the first market research firms, Cherington, Wood, and Roper, and director of the *Fortune Survey,* which was the first national poll to use scientific sampling