1 Data and Decisions

E-Commerce

E-Commerce and mobile commerce have dramatically changed the way the world shops. Online shoppers can buy clothes, food, even cars with the click of a mouse and a digital swipe of their credit card—24 hours a day, 7 days a week. Companies now reach their customers in ways no one could even imagine just a generation ago. Online sales in some sectors, such as clothing and electronics, already account for over 15% of total sales, which is about double what it was five years ago. U.S. adults, on average, currently spend about $1200 a year online, but some projections put that number at nearly $2000 a year by 2016.

The trend in online shopping is worldwide. The amount Australians spend online is expected to grow by $10B in the next five years. The research firm Forrester estimates that global digital retailing is headed toward 15 to 20% of total sales worldwide in the near future.

A few generations ago, many store owners knew their customers well. With that knowledge, they could personalize their suggestions, guessing which items that particular customer might like. Online marketers rely on similar information about customers and potential customers to make decisions. But in today's digital age retailers never meet their customers, so, that information has to be obtained in other ways. How do today's companies know which ads to place on your browser or what order to list the websites from your search? How do marketers know what to advertise and to whom?
The answer is …
Data.

1.1 What Are Data?

"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

"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.

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. Every time you make an online purchase, much more information is actually captured than just the details of the purchase itself. What pages did you search in order to get to your purchase? How much time did you spend looking at each? Companies use this information to make decisions about virtually all phases of their business, from inventory to advertising to website design. 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. The huge capacity of these warehouses has given rise to the term **Big Data** to describe data sets so large that traditional methods of storage and analysis are inadequate. Even though the data amounts are huge, some decisions can be made quickly. 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. But data alone can't help you make better business decisions. You must be able to summarize, model, and understand what the data can tell you. That collection of tools and its associated reasoning is what we call "Statistics."

Statistics plays a role in making sense of our complex world in an astonishing number of ways. 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 or make an informed decision, you need Statistics.

An instructor who wanted to analyze student perceptions of business ethics (a question we'll come back to in a later chapter), couldn't administer a survey to every single university student in the United States. That wouldn't be practical or cost-effective. Instead, she could survey a smaller, representative group of students. Statistics can help us make the leap from a smaller sample of data we have at hand to an understanding of the world at large. Chapter 8 discusses sampling. 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:

- Will the new design of your website increase click-through rates and result in more sales?
- 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?
- Do students around the world perceive issues in business ethics differently?
- 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?

The Essence of Statistics

Our ability to answer questions such as these and make sound business decisions with 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 base 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.

Companies use data to make decisions about nearly every aspect of their business. 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 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 data and statistical analysis to drive business decisions from data whether the purpose is predictive or simply descriptive.

Leading companies are embracing business analytics. 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 data-mining techniques. eBay used analytics to examine its own use of computer resources. Although not obvious to their own technical people, once they crunched the data they found huge inefficiencies. According to Forbes, they were able to "save millions in capital expenditures within the first year."

THE W'S:
WHO
WHAT
WHEN
WHERE
WHY

To begin to make sense of **data**, we first need to understand its context. Newspaper journalists know that the lead paragraph of a good story should establish the "Five W's": *who, what, when, where,* and (if possible) *why*. Often, we add *how* to the list as well. 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.

Table 1.1 shows purchase records from an online music retailer. A table like this is called a **data table**. Each row represents a purchase of a music album. In general, rows of a data table correspond to individual **cases** about which we've recorded some characteristics called **variables.**

Order Number	Name	State/Country	Price	Area Code	Album Download	Gift?	Stock ID	Artist
105-2686834-3759466	Katherine H.	Ohio	5.99	440	Identity	N	B00000I5Y6	James Fortune & Fiya
105-9318443-4200264	Samuel P.	Illinois	9.99	312	Port of Morrow	Y	B000002BK9	The Shins
105-1872500-0198646	Chris G.	Massachusetts	9.99	413	Up All Night	N	B000068ZVQ	Syco Music UK
103-2628345-9238664	Monique D.	Canada	10.99	902	Fallen Empires	N	B000001OAA	Snow Patrol
002-1663369-6638649	Katherine H.	Ohio	11.99	440	Sees the Light	N	B002MXA7Q0	La Sera

Table 1.1 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.

Cases go by a variety of names. 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 1.1, the cases are the individual orders.

The column titles (variable names) tell *what* has been recorded. What does a row of Table 1.1 represent? Be careful. Even if people are involved, the cases may not correspond to people. For example, in Table 1.1, each row represents 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 data table like the one shown in Table 1.1 is sometimes called a **spreadsheet**. Although spreadsheets were designed for accounting, 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.

Spreadsheets are not convenient for really large amounts of data. Amazon has tens of millions of customers and millions of products. But very few customers have purchased more than a few dozen items, so almost all the entries in a spreadsheet of customers by items would be blank—not a very efficient way to store information. For that reason large organizations use relational databases.

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 of all the items sold by the company, including information on price, inventory, and past history, is another relation. Transactions may be held in a third "relation" that references each of the other two relations. Table 1.2 shows a small example.

In statistics, analyses are typically performed on a single relation because all variables must refer to the same cases. But often the data must be retrieved from a relational database, which may require expertise with that software. In the rest of the book, we'll assume that the data have been retrieved and placed in a data table or spreadsheet with variables 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	9/15/13	473859	SC5662	1	UPS 2nd Day	N
T23478924	9/15/13	473859	TH2839	1	UPS 2nd Day	N
T63928934	10/20/13	335682	TH2839	3	UPS Ground	N
T72348299	12/22/13	127389	RS3883	1	Fed Ex Ovnt	Y

Table 1.2 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*); total 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 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

(continued)

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 for 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*).

1.2 Variable Types

When the values of a variable are the names of categories, we call it a **categorical**, or **qualitative, variable.** When the values of a variable are measured numerical quantities with **units**, we call it a **quantitative variable.**

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?"

Categorical, or Quantitative?

When area codes were first introduced all phones had dials. To reduce wear and tear on the dials and to speed calls, the lowest-digit codes (the fastest to dial—those for which the dial spun the least) 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.

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

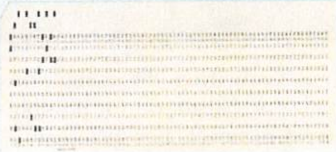
Table 1.3 Some examples of categorical variables.

In a purchase record, price, quantity, and time spent on the website are all quantitative values with units (dollars, count, and seconds). 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 Swazi lilangeni. An essential part of a quantitative variable is its units.

The distinction between categorical and quantitative variables seems clear, but there are reasons to be careful. First, some variables can be considered as either categorical or quantitative, depending on the kind of questions we ask about them. For example, the variable *Age* would be considered quantitative if the responses were numerical and they had units. A doctor would certainly consider *Age* to be quantitative. The units could be years, or for infants, the doctor would want even more precise units, like months, or days. On the other hand, a retailer 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 song download coupon to send you, that might be all the information needed. Then *Age* would be a categorical variable.

Variable Names that Make Sense

A tradition still hangs on in some places 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.



How to classify some variables as categorical or quantitative may seem obvious. But be careful. Area codes may look quantitative, but are really categories. What about ZIP codes? They are categories too, but the numbers do contain information. If you look at a map of the United States with ZIP codes, you'll see that as you move West, the first digit of ZIP codes increases, so treating them as quantitative might make sense for some questions.

Data mining and other statistical analysis programs often must guess whether a variable is categorical or quantitative. When the variable contains symbols other than numbers, the software will correctly treat the variable as categorical, but just because a variable has numbers doesn't mean it is quantitative. Data miners spend much of their time going back through data sets to correctly identify variables as categorical or quantitative. Chapter 2 discusses summaries and displays of categorical variables more fully. Chapter 3 discusses quantitative variables.

Identifiers

Identifier variables are categorical variables that assign a unique identifier code to each individual in the data set. Your student ID number, social security number, and mobile phone number are all identifiers.

Identifier variables are crucial for Big Data because, they make it possible to combine data from different sources, protect confidentiality, and provide unique labels. Your school's grade transcripts are likely in a different relation than your bursar bill records. Your student ID is what links them. Most companies keep such relational databases. The identifiers in Table 1.2 are the *Customer Number*, *Product ID*, and *Transaction Number*.

Other Data Types

Many companies follow up with customers after a service call or sale with an online questionnaire. They 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 average is around 2, but are they *twice* as good? When the values of a categorical variable have an intrinsic order, we can say that the variable is **ordinal**. By contrast, a categorical variable with unordered categories 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. But if we are ordering on purchase volume, it is likely that either Teen or Adult will be the top group.[1]

[1] 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°F 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 quantitative variables that lack a defined zero, 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
2010	10,707
2011	11,700
2012	13,300

Table 1.4 Starbucks's total revenue (in $M) for the years 2002 to 2012.

Cross-Sectional and Time Series Data

The quantitative variable *Total Revenue* in Table 1.4 is an example of a time series. A **time series** is an ordered sequence of values of a single quantitative 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.

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. If we collect data on sales revenue, number of customers, and expenses for last month at *each* Starbucks (more than 21,000 locations as of 2014) 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 5) 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 over time.

1.3 Data Sources: Where, How, and When

In addition to knowing the *who* and *what* of data, 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. Chapter 8 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.

Data can be found in many places. Companies analyze data from their own databases. Some organizations may charge you 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 Pew Research (www.pewresearch.org) offer information on a variety of current social and demographic trends. Data like these are usually collected for different purposes than to answer your particular business question, so you should be cautious when generalizing from data like these. 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. Chapter 16 discusses data mining, which attempts to use Big Data to make hypotheses and draw insights.

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, we often provide a margin note for a new dataset listing some of the W's of the data. When we can, we also 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.

(continued)

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 ($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. One particularly intriguing approach is an interactive software system that uses analogy-based phonics.

Sarah contacted the company that developed this software, RSPT Inc., to obtain the system free of charge for use in her research. RSPT Inc. expressed interest in having her compare its product with other intervention strategies and was quite confident that its 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.

- Identify the ethical dilemma in this scenario.
- What are the undesirable consequences?
- Propose an ethical solution that considers the welfare of all stakeholders.

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, management wished to emphasize the excellent advisory services provided by its 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.

- Identify the ethical dilemma in this scenario.
- What are the undesirable consequences?
- Propose an ethical solution that considers the welfare of all stakeholders.

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

Big Data The collection and analysis of data sets so large and complex that traditional methods typically brought to bear on the problem would be overwhelmed.

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

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.

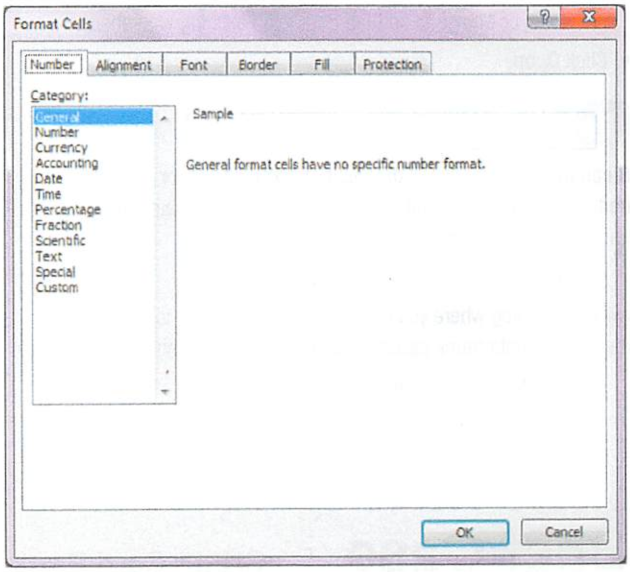
For example, to get your data into a computer statistics package, you need to tell the computer:

- 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.

EXCEL

To open a file containing data in Excel:

- Choose **File > Open**.
- Browse to find the file to open. Data files provided with this text are tab-delimited text files (.txt) or comma-delimited text files (.csv).Excel supports many other file formats.
- You can also copy tables of data from other sources, such as Internet sites, and paste them into an Excel spreadsheet. Excel can recognize the format of many tables copied this way, but this method may not work for some tables.
- When opening a data file, Excel may not recognize the format of the data. If data include dates or other special formats ($, €, ¥, etc.), identify the desired format. Select the cells or columns to reformat and choose **Format > Cell**. Often, the General format is the best option.



General format cells have no specific number format.

JMP

To import a text file, choose:

- **File > Open** and select the file from the dialog. At the bottom of the dialog screen you'll see **Open As:**—be sure to change to **Data (Using Preview)**. This will allow you to specify the delimiter and make sure the variable names are correct. (**JMP** also allows various formats to be imported directly, including .xls files.)

You can also paste a data set in directly (with or without variable names) by selecting:

- **File > New > New Data Table** and then **Edit > Paste (or Paste with Column Names** if you copied the names of the variables as well).

Finally, you can import a data set from a URL directly by selecting:

- **File > Internet Open** and pasting in the address of the web site. JMP will attempt to find data on the page. It may take a few tries and some edits to get the data set in correctly.

MINITAB

To import a text or Excel file, choose:

- **File > Open Worksheet**. From **Files of type,** choose **Text (*.txt)** or **Excel (*.xls; *xlsx).**

- Browse to find and select the file.
- In the lower right corner of the dialog, choose **Open** to open the data file alone, or **Merge** to add the data to an existing worksheet.
- Click **Open**.

R

R can import many types of files, but text files (tab or comma delimited) are easiest. If the file is tab delimited and contains the variable names in the first row, then:

```
> mydata = read.delim(file.choose())
```

will give a dialog where you can pick the file you want to import. It will then be in a data frame called mydata. If the file is comma delimited, use:

```
>mydata = read.csv(file.choose())
```

Comments

(RStudio provides an interactive dialog that may be easier to use). For other options including the case that the file does not contain variable names, consult **R** help.

SPSS

To import a text file, choose:

- **File > Open > Data.** Under "Files of type", choose **Text (*.txt,*.dat).** Select the file you want to import. Click **Open.**
- A window will open called **Text Import Wizard.** Follow the steps, depending on the type of file you want to import.

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 1.1

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	Size
41340053	Greenfield Manor	12859	1.00	1967	$1,00,400	960
4128001474	Fort Amherst	12801	0.09	1961	$1,32,500	906
412800344	Dublin	12309	1.65	1993	$1,40,000	1620
4128001552	Granite Springs	10598	0.33	1969	$67,100	900
412800352	Arcady	10562	2.29	1955	$1,90,000	1224
413400322	Ormsbee	12859	9.13	1997	$1,26,900	1056

2. A local bookstore is keeping a database of its customers to find out more about their spending habits 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 1.2

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 1.3

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 (if they are not provided, give some possible units in which they might be measured). 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.

7. The news. Find a newspaper or magazine article in which some data are reported (e.g., see *The Wall Street Journal*, *Financial Times*, *Business Week*, or *Fortune*). For the data discussed in the article, answer the questions above. Include a copy of the article with your report.

8. The Internet. Find an Internet site on which some data are reported. For the data found on the site, answer as many of the questions above as you can. Include a copy of the URL with your report.

9. Survey. An automobile manufacturer wants to know what college students think about electric vehicles. They ask you to conduct a survey that asks students, "Do you think there will be more electric or gasoline powered vehicles on the road in 2025?" and "How likely are you to buy an electric vehicle in the next 10 years?" (scale of 1 = not at all likely to 5 = very likely).

10. Your survey. Think of a question that you'd like to know the answer to that might be answered with a survey. What are the questions? Identify the variables and answer the questions above.

11. World databank. The World Bank provides economic data on most of the world's countries at their website (databank.worldbank.org/data/home.aspx). Select 5 indicators that they provide and answer the questions above for these variables.

12. Arby's menu. A listing posted by the Arby's restaurant chain gives, for each of the sandwiches it sells, the type of meat in the sandwich, number of calories, and serving size in ounces. The data might be used to assess the nutritional value of the different sandwiches.

13. MBA admissions. A school in the northeastern United States is concerned with the recent drop in female students in its MBA program. It decides to collect data from the admissions office on each applicant, including: sex of each applicant, age of each applicant, whether or not they were accepted, whether or not they attended, and the reason for not attending (if they did not attend). The school hopes to find commonalities among the female accepted students who have decided not to attend the business program.

14. MBA admissions II. An internationally recognized MBA program outside of Paris intends to also track the GPA of the MBA students and compares MBA performance to standardized test scores over a six-year period (2009–2014).

15. Pharmaceutical firm. Scientists at a major pharmaceutical firm conducted an experiment to study the effectiveness of an herbal compound to treat the common cold. They exposed volunteers to a cold virus, then gave them either the herbal compound or a sugar solution known to have no effect on colds. Several days later they assessed each patient's condition using a cold severity scale ranging from 0–5. They found no evidence of the benefits of the compound.

16. Start-up company. A start-up company is building a database of customers and sales information. For each customer, it records name, ID number, region of the country (1 = East, 2 = South, 3 = Midwest, 4 = West), date of last purchase, amount of purchase, and item purchased.

17. Vineyards. Business analysts hoping to provide information helpful to grape growers sent out a questionnaire to a sample of growers requesting these data about vineyards: size, number of years in existence, state, varieties of grapes grown, average case price, gross sales, and percent profit.

18. Spectrem Group polls. Spectrem Group (www.spectrem.com) provides services for the affluent and retirement markets. In a recent survey, they found that millionaires tend to prefer dogs to cats from a question asking them to list the pets they own. They also found that senior executives are more likely to buy treats and toys for their pet than regular investors by asking, "What services and products do you buy for your pet?"

19. EPA. The Environmental Protection Agency (EPA) tracks fuel economy of automobiles. Among the data EPA analysts collect from the manufacturer are the manufacturer (Ford, Toyota, etc.), vehicle type (car, SUV, etc.), weight, horsepower, and gas mileage (mpg) for city and highway driving.

20. Consumer Reports. In 2013, Consumer Reports published an article comparing smart phones. It listed 46 phones, giving brand, price, display size, operating system (Android, iOS, or Windows Phone), camera image size (megapixels), and whether it had a memory card slot.

21. Zagat. Zagat.com provides ratings from customer experiences on restaurants. For each restaurant, the percentage of customers that liked it, the average cost and ratings of the food, decor, and service (all on a 30-point scale) are reported.

22. L.L. Bean. L.L. Bean is a large U.S. retailer that depends heavily on its catalog sales. It collects data internally and tracks the number of catalogs mailed out, the number of square inches in each catalog, and the sales ($ thousands) in the four weeks following each mailing. The company is interested in learning more about the relationship (if any) among the timing and space of their catalogs and their sales.

23. Stock market. An online survey of students in a large MBA Statistics class at a business school in the northeastern United States asked them to report their total personal investment in the stock market ($), total number of different stocks currently held, total invested in mutual funds ($), and the name of each mutual fund in which they have invested. The data were used in the aggregate for classroom illustrations.

24. Theme park sites. A study on the potential for developing theme parks in various locations throughout Europe in 2013 collects the following information: the country where the proposed site is located, estimated cost to acquire site, size of population within a one-hour drive of the site, size of the site, and availability of mass transportation within five minutes of the site. The data will be used to present to prospective developers.

25. Indy 2014. The 2.5-mile Indianapolis Motor Speedway has been the home to a race on Memorial Day nearly every year since 1911. Even during the first race there were controversies. Ralph Mulford was given the checkered flag first but took three extra laps just to make sure he'd completed 500 miles. When he finished, another driver, Ray Harroun, was being presented with the winner's trophy, and Mulford's protests were ignored. Harroun averaged 74.6 mph for the 500 miles. Here are the data for the first few and six recent Indianapolis 500 races.

Year	Winner	Chassis	Engine	Time (hrs)	Speed (mph)	Car #
1911	Ray Harroun	Marmon	Marmon	6.7022	74.602	32
1912	Joe Dawson	National	National	6.3517	78.719	8
1913	Jules Goux	Peugeot	Peugeot	6.5848	75.933	16
...		...				
...		...				
2009	Hélio Castroneves	Dallara	Honda	3.3262	150.318	3
2010	Dario Franchitti	Dallara	Honda	3.0936	161.623	10
2011	Dan Wheldon	Dallara	Chevrolet	2.9366	170.265	98
2012	Dario Franchitti	Dallara	Honda	2.9809	167.734	50
2013	Tony Kanaan	Dallara	Chevrolet	2.6676	187.433	12
2014	Ryan Hunter-Reay	Dallara	Honda	2.6801	186.563	19

T **26.** Kentucky Derby 2014. The Kentucky Derby is a horse race that has been run every year since 1875 at Churchill Downs, Louisville, Kentucky. The race started as a 1.5-mile race, but in 1896 it was shortened to 1.25 miles because experts felt that three-year-old horses shouldn't run such a long race that early in the season. (It has been run in May every year but one—1901—when it took place on April 29.) The table at the bottom of the page shows the data for the first few and a few recent races. http://www.kentuckyderby.ag/kentuckyderby-results.php and http://horseracing.about.com/od/history/l/blderbywin.htm

Year	Winner	Jockey	Duration (seconds)	Track Condition
1875	Aristides	O. Lewis	157.75	Fast
1876	Vagrant	B. Swim	158.25	Fast
1877	Baden-Baden	W. Walker	158	Fast
1878	Day Star	J. Carter	157.25	Dusty
1879	Lord Murphy	C. Shauer	157	Fast
...				
2010	Super Saver	Calvin Borel	124.45	Fast
2011	Animal Kingdom	John R. Velazquez	122.04	Fast
2012	I'll Have Another	Mario Gutierrez	121.83	Fast
2013	Orb	J. Rosario	122.89	Sloppy
2014	California Chrome	Victor Espinoza	123.66	Fast

When you organize data in a spreadsheet, it is important to lay it out as a data table. For each of these examples in Exercises 27 to 30, show how you would lay out these data. Indicate the headings of columns and what would be found in each row.

27. Mortgages. For a study of mortgage loan performance: amount of the loan, the name of the borrower.

28. Employee performance. Data collected to determine performance-based bonuses: employee ID, average contract 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 time-series or cross-sectional.

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 2014.

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