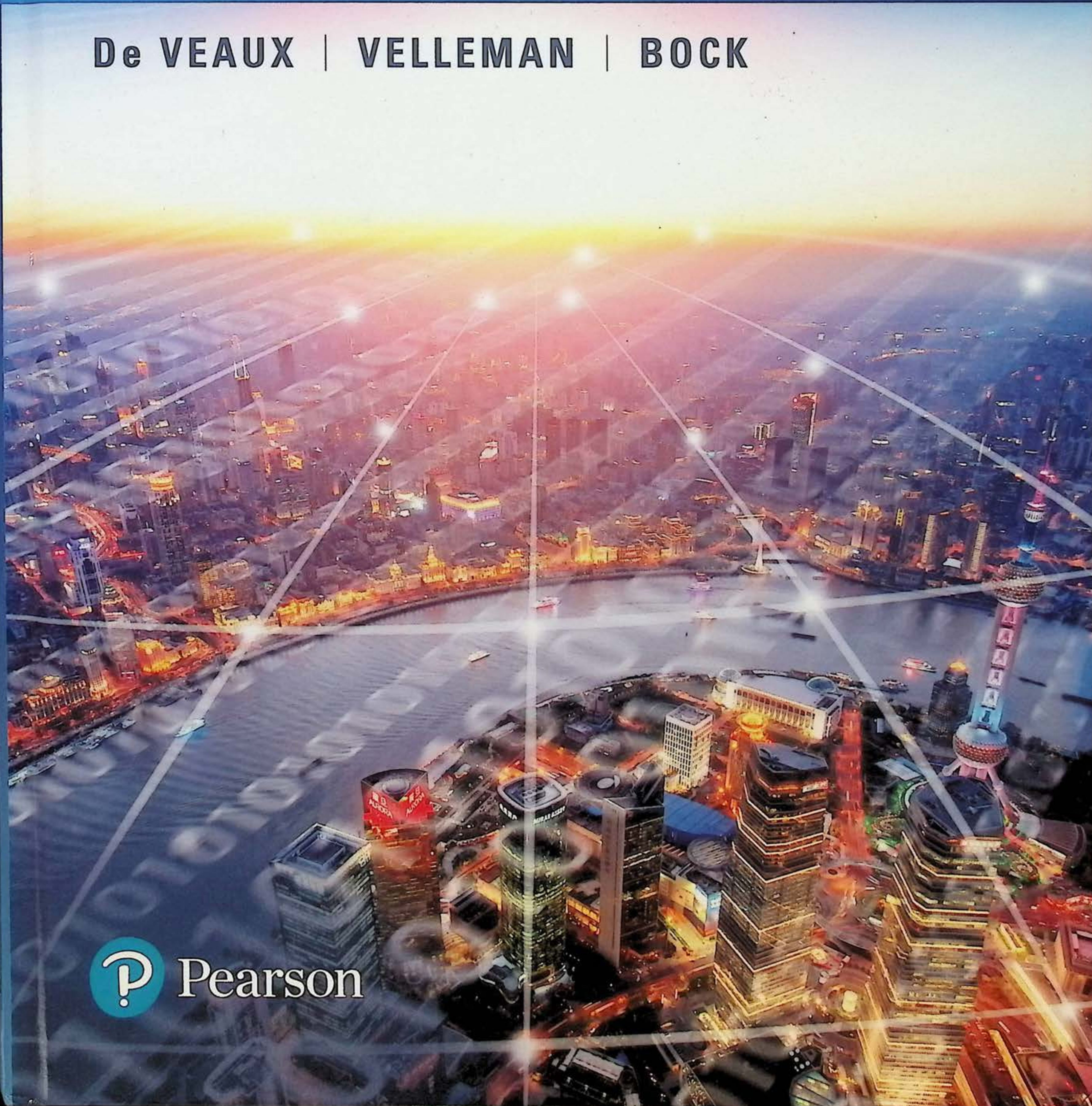
Intro STATS

FIFTH EDITION

De VEAUX | VELLEMAN | BOCK



Pearson

TABLE OF CONTENTS

PART I Exploring and Understanding Data

PART II Exploring Relationships Between Variables

*Indicates optional sections.

PART III Gathering Data

PART IV From the Data at Hand to the World at Large

1

Stats Starts Here[1]

WHERE ARE WE GOING?

Statistics gets no respect. People say things like "You can prove anything with statistics." People will write off a claim based on data as "just a statistical trick." And statistics courses don't have the reputation of being students' first choice for a fun elective.

But statistics *is* fun. That's probably not what you heard on the street, but it's true. Statistics is the science of learning from data. A little practice thinking statistically is all it takes to start seeing the world more clearly and accurately.

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

1.1 What Is Statistics?

People around the world have one thing in common—they all want to figure out what's going on. You'd think with the amount of information available to everyone today this would be an easy task, but actually, as the amount of information grows, so does our need to understand what it can tell us.

At the base of all this information, on the Internet and all around us, are data. We'll talk about data in more detail in the next section, but for now, think of **data** as any collection of numbers, characters, images, or other items that provide information about something. What sense can we make of all this data? You certainly can't make a coherent picture from random pieces of information. Whenever there are data and a need for understanding the world, you'll find statistics.

> "But where shall I begin?" asked Alice. "Begin at the beginning," the King said gravely, "and go on till you come to the end: then stop."
>
> —Lewis Carroll, Alice's Adventures in Wonderland

This book will help you develop the skills you need to understand and communicate the knowledge that can be learned from data. By thinking clearly about the question you're trying to answer and learning the statistical tools to show what the data are saying, you'll acquire the skills to tell clearly what it all means. Our job is to help you make sense of the concepts and methods of statistics and to turn it into a powerful, effective approach to understanding the world through data.

[1] We were thinking of calling this chapter "Introduction" but nobody reads the introduction, and we wanted you to read this. We feel safe admitting this down here in the footnotes because nobody reads footnotes either.

STATISTICS SHOW FAR FEWER SKATEBOARDING INJURIES IN THE WINTER, WHILE SLEDDING INJURIES ARE MORE COMMON.
CLEARLY, I'M SAFER GOING DOWN POPSICLE PRECIPICE ON MY SKATEBOARD.
FOUR OUT OF FIVE PEOPLE WHO DON'T QUESTION STATISTICS END UP IN JOBS INVOLVING NECKTIES.
REALLY?
©2003 Jef Mallett/Distributed by United Feature Syndicate, Inc.
www.comics.com jefmallett@yahoo.com
MALLETT
2-24

FRAZZ © 2003 Jef Mallett. Distributed by Andrews McMeel Syndication. Reprinted with permission. All rights reserved.

> "Data is king at Amazon. Clickstream and purchase data are the crown jewels at Amazon. They help us build features to personalize the Web site experience."
>
> —Ronny Kohavi, former Director of Data Mining and Personalization, Amazon.com

Q: What is statistics?

A: Statistics is a way of reasoning, along with a collection of tools and methods, designed to help us understand the world.

Q: What are statistics?

A: Statistics (plural) are particular calculations made from data.

Q: So what is data?

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

Q: OK, OK, so what are data?

A: Data are values along with their context.

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

We say, "Don't be a datum."

Data vary. Ask different people the same question and you'll get a variety of answers. Statistics helps us to make sense of the world described by our data by seeing past the underlying variation to find patterns and relationships. This book will teach you skills to help with this task and ways of thinking about variation that are the foundation of sound reasoning about data.

Consider the following:

- If you have a Facebook account, you have probably noticed that the ads you see online tend to match your interests and activities. Coincidence? Hardly. According to *The Wall Street Journal* (10/18/2010),[2] much of your personal information has probably been sold to marketing or tracking companies. Why would Facebook give you a free account and let you upload as much as you want to its site? Because your data are valuable! Using your Facebook profile, a company might build a profile of your interests and activities: what movies and sports you like; your age, sex, education level, and hobbies; where you live; and, of course, who your friends are and what *they* like. From Facebook's point of view, your data are a potential gold mine. Gold ore in the ground is neither very useful nor pretty. But with skill, it can be turned into something both beautiful and valuable. What we're going to talk about in this book is how you can mine your own data and learn valuable insights about the world.
- Americans spend an average of 4.9 hours per day on their smartphones. Trillions of text messages are sent each year.[3] Some of these messages are sent or read while the sender or the receiver is driving. How dangerous is texting while driving?

  How can we study the effect of texting while driving? One way is to measure reaction times of drivers faced with an unexpected event while driving and texting. Researchers at the University of Utah tested drivers on simulators that could present emergency situations. They compared reaction times of sober drivers, drunk drivers, and texting drivers.[4] The results were striking. The texting drivers actually responded more slowly and were more dangerous than drivers who were above the legal limit for alcohol.

In this book, you'll learn how to design and analyze experiments like this. You'll learn how to interpret data and to communicate the message you see to others. You'll also learn how to spot deficiencies and weaknesses in conclusions drawn by others that you see in newspapers and on the Internet every day. Statistics can help you become a more informed citizen by giving you the tools to understand, question, and interpret data.

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

[3] http://informatemi.com/blog/?p=133

[4] "Text Messaging During Simulated Driving," Drews, F. A., et al., Human Factors: hfs.sagepub.com/content/51/5/762

1.2 Data

STATISTICS IS ABOUT ...

- Variation: Data vary because we don't see everything, and even what we do see, we measure imperfectly.
- Learning from data: We hope to learn about the world as best we can from the limited, imperfect data we have.
- Making intelligent decisions: The better we understand the world, the wiser our decisions will be.

Amazon.com opened for business in July 1995, billing itself as "Earth's Biggest Bookstore." By 1997, Amazon had a catalog of more than 2.5 million book titles and had sold books to more than 1.5 million customers in 150 countries. In 2016, the company's sales reached almost $136 billion (more than 25% over the previous year). Amazon has sold a wide variety of merchandise, including a $400,000 necklace, yak cheese from Tibet, and the largest book in the world. How did Amazon become so successful and how can it keep track of so many customers and such a wide variety of products? The answer to both questions is *data*.

But what are data? Think about it for a minute. What exactly *do* we mean by "data"? You might think that data have to be numbers, but data can be text, pictures, web pages, and even audio and video. If you can sense it, you can measure it. Data are now being collected automatically at such a rate that IBM estimates that "90% of the data in the world today has been created in the last two years alone."[5]

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

B0000010AA	0.99	Chris G.	902	105-2686834-3759466	1.99	0.99	Illinois
Los Angeles	Samuel R.	Ohio	N	B000068ZVQ	Amsterdam	New York, New York	Katherine H.
Katherine H.	002-1663369-6638649	Beverly Hills	N	N	103-2628345-9238664	0.99	Massachusetts
312	Monique D.	105-9318443-4200264	413	B00000I5Y6	440	B000002BK9	0.99
Canada	Detroit	440	105-1372500-0198646	N	B002MXA7Q0	Ohio	Y

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

Order Number	Name	State/Country	Price	Area Code	Download	Gift?	ASIN	Artist
105-2686834-3759466	Katherine H.	Ohio	0.99	440	Amsterdam	N	B0000015Y6	Cold Play
105-9318443-4200264	Samuel R	Illinois	1.99	312	Detroit	Y	B000002BK9	Red Hot Chili Peppers
105-1372500-0198646	Chris G.	Massachusetts	0.99	413	New York, New York	N	B000068ZVQ	Frank Sinatra
103-2628345-9238664	Monique D.	Canada	0.99	902	Los Angeles	N	B0000010AA	Blink 182
002-1663369-6638649	Katherine H.	Ohio	0.99	440	Beverly Hills	N	B002MXA7Q0	Weezer

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

[5] http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html

What information would provide a **context**? Newspaper journalists know that the lead paragraph of a good story should establish the "Five W's": *who, what, when, where,* and (if possible) *why*. Often, we add *how* to the list as well. The answers to the first two questions are essential. If we don't know *what* values are measured and *who* those values are measured on, the values are meaningless.

Who and What

In general, the rows of a data table correspond to individual **cases** about *whom* (or about which, if they're not people) we record some characteristics. Cases go by different names, depending on the situation.

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

DATA BEATS INTUITION

Amazon monitors and updates its website to better serve customers and maximize sales. To decide which changes to make, analysts experiment with new designs, offers, recommendations, and links. Statisticians want to know how long you'll spend browsing the site and whether you'll follow the links or purchase the suggested items. As Ronny Kohavi, former director of Data Mining and Personalization for Amazon, said, "Data trumps intuition. Instead of using our intuition, we experiment on the live site and let our customers tell us what works for them."

Look at all the columns to see exactly what each row refers to. Here the cases are different purchase records. You might have thought that each customer was a case, but notice that, for example, Katherine H. appears twice, in both the first and the last row. A common place to find out exactly what each row refers to is the leftmost column. That value often identifies the cases, in this example, it's the order number. If you collect the data yourself, you'll know what the cases are. But, often, you'll be looking at data that someone else collected and you'll have to ask or figure that out yourself.

Often the cases are a **sample** from some larger **population** that we'd like to understand. Amazon doesn't care about just these customers; it wants to understand the buying patterns of *all* its customers, and, generalizing further, it wants to know how to attract other Internet users who may not have made a purchase from Amazon's site. To be able to generalize from the sample of cases to the larger population, we'll want the sample to be *representative* of that population—a kind of snapshot image of the larger world.

We must know *who* and *what* to analyze data. Without knowing these two, we don't have enough information to start. Of course, we'd always like to know more. The more we know about the data, the more we'll understand about the world. If possible, we'd like to know the *when* and *where* of data as well. Values recorded in 1803 may mean something different than similar values recorded last year. Values measured in Tanzania may differ in meaning from similar measurements made in Mexico. And knowing *why* the data were collected can tell us much about its reliability and quality.

How the Data Are Collected

How the data are collected can make the difference between insight and nonsense. As we'll see later, data that come from a voluntary survey on the Internet are almost always worthless. One primary concern of statistics, to be discussed in Part III, is the design of sound methods for collecting data. Throughout this book, whenever we introduce data, we'll provide a margin note listing the W's (and H) of the data. Identifying the W's is a habit we recommend.

The first step of any data analysis is to know what you are trying to accomplish and what you want to know. To help you use statistics to understand the world and make decisions, we'll lead you through the entire process of *thinking* about the problem, *showing* what you've found, and *telling* others what you've learned. Every guided example in this book is broken into these three steps: *Think*, *Show*, and *Tell*. Identifying the problem and the *who* and *what* of the data is a key part of the *Think* step of any analysis. Make sure you know these before you proceed to *Show* or *Tell* anything about the data.

EXAMPLE 1.1

Identifying the *Who*

In 2015, *Consumer Reports* published an evaluation of 126 tablets from a variety of manufacturers.

QUESTION: Describe the population of interest, the sample, and the *Who* of the study.

ANSWER: The magazine is interested in the performance of tablets currently offered for sale. It tested a sample of 126 tablets, which are the *Who* for these data. Each tablet selected represents all similar tablets offered by that manufacturer.

1.3 Variables

The characteristics recorded about each individual are called **variables**. They are usually found as the columns of a data table with a name in the header that identifies what has been recorded. In the Amazon data table we find the variables *Order Number, Name, State/Country, Price*, and so on.

Categorical Variables

Some variables just tell us what group or category each individual belongs to. Are you male or female? Pierced or not? We call variables like these **categorical**, or **qualitative, variables**. (You may also see them called **nominal variables** because they name categories.) Some variables are clearly categorical, like the variable *State/Country*. Its values are text and those values tell us what category the particular case falls into. But numerals are often used to label categories, so categorical variable values can also be numerals. For example, Amazon collects telephone area codes that *categorize* each phone number into a geographical region. So area code is considered a categorical variable even though it has numeric values. (But see the story in the following box.)

AREA CODES—NUMBERS OR CATEGORIES?

The *What* and *Why* of area codes are not as simple as they may first seem. When area codes were first introduced, AT&T was still the source of all telephone equipment, and phones had dials.

To reduce wear and tear on the dials, the area codes with the lowest digits (for which the dial would have to spin least) were assigned to the most populous regions—those with the most phone numbers and thus the area codes most likely to be dialed. New York City was assigned 212, Chicago 312, and Los Angeles 213, but rural upstate New York was given 607, Joliet was 815, and San Diego 619. For that reason, at one time the numerical value of an area code could be used to guess something about the population of its region. Since the advent of push-button phones, area codes have finally become just categories.

Descriptive responses to questions are often categories. For example, the responses to the questions "Who is your cell phone provider?" and "What is your marital status?" yield categorical values. When Amazon considers a special offer of free shipping to customers, it might first analyze how purchases have been shipped in the recent past. Amazon might start by counting the number of purchases shipped in each category: ground transportation, second-day air, and next-day air. Counting is a natural way to summarize a categorical variable such as *Shipping Method*. Chapter 2 discusses summaries and displays of categorical variables more fully.

> Far too many scientists have only a shaky grasp of the statistical techniques they are using. They employ them as an amateur chef employs a cookbook, believing the recipes will work without understanding why. A more *cordon bleu* attitude . . . might lead to fewer statistical soufflés failing to rise.
>
> —The Economist, June 3, 2004, "Sloppy stats shame science"

THERE'S A WORLD OF DATA ON THE INTERNET

These days, one of the richest sources of data is the Internet. With a bit of practice, you can learn to find data on almost any subject. Many of the data sets we use in this book were found in this way. The Internet has both advantages and disadvantages as a source of data. Among the advantages are the fact that often you'll be able to find even more current data than those we present. The disadvantage is that references to Internet addresses can "break" as sites evolve, move, and die.

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

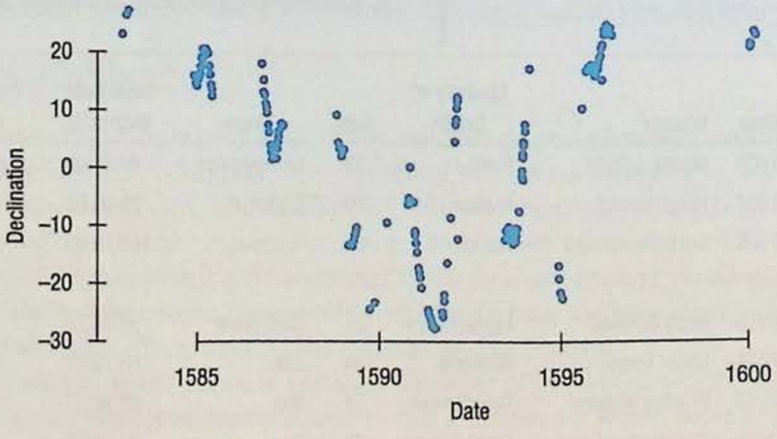
Some words of caution, though: Data found on Internet sites may not be formatted in the best way for use in statistics software. Although you may see a data table in standard form, an attempt to copy the data may leave you with a single column of values. You may have to work in your favorite statistics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

1.4 Models

What is a **model** for data? Models are summaries and simplifications of data that help our understanding in many ways. We'll encounter all sorts of models throughout the book. A model is a simplification of reality that gives us information that we can learn from and use, even though it doesn't represent reality exactly. A model of an airplane in a wind tunnel can give insights about the aerodynamics and flight performance of the plane even though it doesn't show every rivet.[9] In fact, it's precisely because a model is a simplification that we learn from it. Without making models for how data vary, we'd be limited to reporting only what the data we have at hand says. To have an impact on science and society we'll have to generalize those findings to the world at large.

Kepler's laws describing the motion of planets are a great example of a model for data. Using astronomical observations of Tycho Brahe, Kepler saw through the small anomalies in the measurements and came up with three simple "laws"—or models for how the planets move. Here are Brahe's observations on the declination (angle of tilt to the sun) of Mars over a twenty-year period just before 1600:

Figure 1.1
A plot of declination against time shows some patterns. There are many missing observations. Can you see the model that Kepler came up with from these data?



Here, using modern statistical methods is a plot of the model predictions from the data:

[9] Or tell you what movies you might see on the flight.

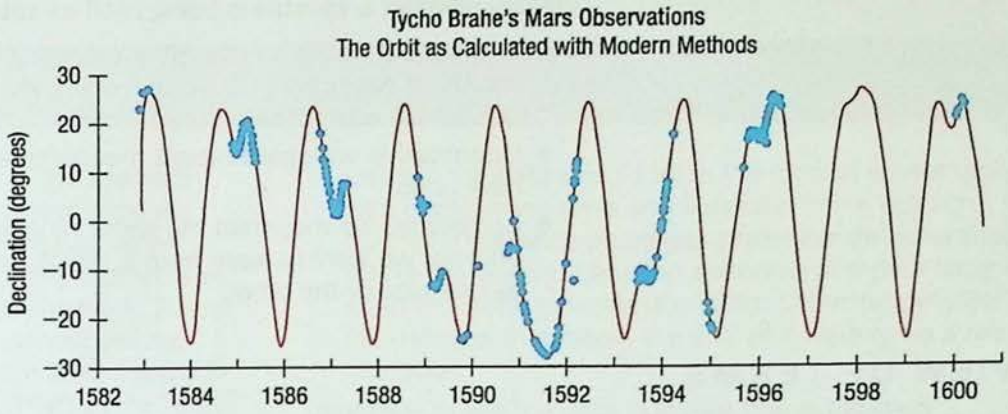


Figure 1.2
The model that Kepler proposed filled in many of the missing points and made the pattern much clearer.

Later, after Newton laid out the physics of gravity, it could be shown that the laws follow from other principles, but Kepler derived the models from data. We may not be able to come up with models as profound as Kepler's, but we'll use models throughout the book. We'll see examples of models as early as Chapter 5 and then put them to use more thoroughly later in the book when we discuss inference.

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.

CHAPTER REVIEW

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.

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.

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.

REVIEW OF TERMS

The key terms are in chapter order so you can use this list to review the material in the chapter.

Data Recorded values, whether numbers or labels, together with their context (p. 1).

Data table An arrangement of data in which each row represents a case and each column represents a variable (p. 3).

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 (p. 4).

Case An individual about whom or which we have data (p. 4).

Respondent Someone who answers, or responds to, a survey (p. 4).

Subject A human experimental unit. Also called a participant (p. 4).

Participant A human experimental unit. Also called a subject (p. 4).

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 (p. 4).

Record Information about an individual in a database (p. 4).

Sample A subset of a population, examined in hope of learning about the population (p. 4).

Population The entire group of individuals or instances about whom we hope to learn (p. 4).

Variable A variable holds information about the same characteristic for many cases (p. 5).

Categorical (or qualitative) variable A variable that names categories with words or numerals (p. 5).

Nominal variable The term "nominal" can be applied to a variable whose values are used only to name categories (p. 5).

Quantitative variable A variable in which the numbers are values of measured quantities with units (p. 6).

Units A quantity or amount adopted as a standard of measurement, such as dollars, hours, or grams (p. 6).

Identifier variable A categorical variable that records a unique value for each case, used to name or identify it (p. 6).

Ordinal variable The term "ordinal" can be applied to a variable whose categorical values possess some kind of order (p. 7).

Model A description or representation, in mathematical and statistical terms, of the behavior of a phenomenon based on data (p. 8).

TECH SUPPORT

Entering Data

These days, nobody does statistics by hand. We use technology: a programmable calculator or a statistics program on a computer. Professionals all use a *statistics package* designed for the purpose. We will provide many examples of results from a statistics package throughout the book. Rather than choosing one in particular, we'll offer generic results that look like those produced by all the major statistics packages but don't exactly match any of them. Then, in the Tech Support section at the end of each chapter, we'll provide hints for getting started on several of the major packages.

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 with cases in the rows and variables in the columns. Most computer statistics packages prefer the *delimiter* that marks the division between elements of a data table to be a tab character (comma is another common delimiter) 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 is often used to help organize, manipulate, and prepare data for other software packages. Many of the other packages take Excel files as inputs. Alternatively, you can copy a data table from Excel and Paste it into many packages, or export Excel spreadsheets as tab delimited (.txt) or comma delimited files (.csv), which can be easily shared and imported into other programs. All data files provided with this text are in tab-delimited text (.txt) format.

EXCEL

To open a file containing data in Excel:

▶ Choose **File** > **Open**.

▶ Browse to find the file to open. Excel supports many file formats.

▶ Other programs can import data from a variety of file formats, but all can read both tab delimited (.txt) and comma delimited (.csv) text files.

▶ 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.

▶ 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 for data you plan to move to a statistics package.

DATA DESK

To read data into Data Desk:

▶ Click the **Open File** icon or choose **File** > **Open**. The dialog lets you specify variable names (or take them from the first row of the data), the delimiter, or how to read formatted data.

▶ **File** > **Import** works the same way, but instead of starting a new data file, it adds the data in the file to the current data file. Data Desk can work with multiple data tables in the same file.

▶ If the data are already in another program, such as, for example, a spreadsheet, **Copy** the data table (including the column headings). In Data Desk choose **Edit** > **Paste** variables. There is no need to create variables first; Data Desk does that automatically. You'll see the same dialog as for Open and Import.

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 website. 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.

STATCRUNCH

Statcrunch offers several ways to enter data. Click **MyStatCrunch > My Data**. Click a dataset to analyze the data or edit its properties.

Click a data set link to analyze the data or edit its properties to import a new data set.

- Choose **Select a file on my computer,**
- Enter the URL of a file,
- Paste data into a form, or
- Type or paste data into a blank data table.

For the "select a file on my computer" option, Statcrunch offers a choice of space, comma, tab, or semicolon delimiters. You may also choose to use the first line as the names of the variables.

After making your choices, select the **Load File** button at the bottom of the screen.

EXERCISES

SECTION 1.1

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

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

3. Parking lots Sensors in parking lots are able to detect and communicate when spaces are filled in a large covered parking garage next to an urban shopping mall. How might the owners of the parking garage use this information both to attract customers and to help the store owners in the mall make business plans?

4. Satellites and global climate change Satellites send back nearly continuous data on the earth's land masses, oceans, and atmosphere from space. How might researchers use this information in both the short and long term to help study changes in the earth's climate?

SECTION 1.2

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

6. Nobel laureates The website www.nobelprize.org allows you to look up all the Nobel prizes awarded in any year. The data are not listed in a table. Rather you drag a slider to the year and see a list of the awardees for that year. Describe the *Who* in this scenario.

7. Health records The National Center for Health Statistics (NCHS) conducts an extensive survey consisting of an interview and medical examination with a representative sample of about 5000 people a year. The interview includes demographic, socioeconomic, dietary, and other health-related questions. The examination "consists of medical, dental, and physiological measurements, as well as laboratory tests administered by highly trained medical personnel" (www.cdc.gov/nchs/nhanes/about_nhanes.htm). Describe the sample, the population, the *Who* and the *What* of this study.

8. Facebook. Facebook uploads more than 350 million photos every day onto its servers. For this collection, describe the *Who* and the *What*.

SECTION 1.3

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

10. ZIP codes The U.S. Postal Service uses five-digit ZIP codes to identify locations to assist in delivering mail.
    a) In what sense are ZIP codes categorical?
    b) Is there any ordinal sense to ZIP codes? In other words, does a higher ZIP code tell you anything about a location compared to a lower ZIP code?

11. Voters A February 2010 Gallup Poll question asked, "In politics, as of today, do you consider yourself a Republican, a Democrat, or an Independent?" The possible responses were "Democrat," "Republican," "Independent," "Other," and "No Response." What kind of variable is the response?

12. Job hunting A June 2011 Gallup Poll asked Americans, "Thinking about the job situation in America today, would you say that it is now a good time or a bad time to find a quality job?" The choices were "Good time" or "Bad time." What kind of variable is the response?

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

14. Stress A medical researcher measures the increase in heart rate of patients who are taking a stress test. What kind of variable is the researcher studying?

SECTION 1.4

15. Voting and elections Pollsters are interested in predicting the outcome of elections. Give an example of how they might model whether someone is likely to vote.

16. Weather Meteorologists utilize sophisticated models to predict the weather up to ten days in advance. Give an example of how they might assess their models.

17. The news Find a newspaper or magazine article in which some data are reported. For the data discussed in the article, identify as many of the W's as you can. Include a copy of the article with your report.

18. The Internet Find an Internet source that reports on a study and describes the data. Print out the description and identify as many of the W's as you can.

***(Exercises 19–26)** For each description of data, identify Who and What were investigated and the Population of interest.*

19. **Gaydar** A study conducted by a team of American and Canadian researchers found that during ovulation, a woman can tell whether a man is gay or straight by looking at his face. To explore the subject, the authors conducted three investigations, the first of which involved 40 undergraduate women who were asked to guess the sexual orientation of 80 men based on photos of their face. Half of the men were gay, and the other half were straight. All held similar expressions in the photos or were deemed to be equally attractive. None of the women were using any contraceptive drugs at the time of the test. The result: the closer a woman was to her peak ovulation, the more accurate her guess. (health.usnews.com/health-news/family-health/brain-and-behavior/articles/2011/06/27/ovulation-seems-to-aid-womens-gaydar)

20. **Hula-hoops** The hula-hoop, a popular children's toy in the 1950s, has gained popularity as an exercise in recent years. But does it work? To answer this question, the American Council on Exercise conducted a study to evaluate the cardio and calorie-burning benefits of "hooping." Researchers recorded heart rate and oxygen consumption of participants, as well as their individual ratings of perceived exertion, at regular intervals during a 30-minute workout. (www.acefitness.org/certifiednewsarticle/1094/)

21. **Bicycle safety** Ian Walker, a psychologist at the University of Bath, wondered whether drivers treat bicycle riders differently when they wear helmets. He rigged his bicycle with an ultrasonic sensor that could measure how close each car was that passed him. He then rode on alternating days with and without a helmet. Out of 2500 cars passing him, he found that when he wore his helmet, motorists passed 3.35 inches closer to him, on average, than when his head was bare. (Source: *NY Times*, Dec. 10, 2006)

22. **Investments** Some companies offer 401(k) retirement plans to employees, permitting them to shift part of their before-tax salaries into investments such as mutual funds. Employers typically match 50% of the employees' contribution up to about 6% of salary. One company, concerned with what it believed was a low employee participation rate in its 401(k) plan, sampled 30 other companies with similar plans and asked for their 401(k) participation rates.

23. **Honesty** Coffee stations in offices often just ask users to leave money in a tray to pay for their coffee, but many people cheat. Researchers at Newcastle University alternately taped two posters over the coffee station. During one week, it was a picture of flowers; during the other, it was a pair of staring eyes. They found that the average contribution was significantly higher when the eyes poster was up than when the flowers were there. Apparently, the mere feeling of being watched—even by eyes that were not real—was enough to encourage people to behave more honestly. (Source: *NY Times*, Dec. 10, 2006)

24. **Blindness** A study begun in 2011 examines the use of stem cells in treating two forms of blindness, Stargardt's disease and dry age-related macular degeneration. Each of the 24 patients entered one of two separate trials in which embryonic stem cells were to be used to treat the condition. (www.blindness.org/index.php?view=article&id=2514:stem-cell-clinical-trial-for-stargardt-disease-set-to-begin-&option=com_content&Itemid=122)

25. **Not-so-diet soda** A look at 474 participants in the San Antonio Longitudinal Study of Aging found that participants who drank two or more diet sodas a day "experienced waist size increases six times greater than those of people who didn't drink diet soda." (*J Am Geriatr Soc.* 2015 Apr;63(4):708–15. doi: 10.1111/jgs.13376. Epub 2015 Mar 17.)

26. **Molten iron** The Cleveland Casting Plant is a large, highly automated producer of gray and nodular iron automotive castings for Ford Motor Company. The company is interested in keeping the pouring temperature of the molten iron (in degrees Fahrenheit) close to the specified value of 2550 degrees. Cleveland Casting measured the pouring temperature for 10 randomly selected crankshafts.

***(Exercises 27–40)** For each description of data, identify the W's, name the variables, specify for each variable whether its use indicates that 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).*

27. **Weighing bears** Because of the difficulty of weighing a bear in the woods, researchers caught and measured 54 bears, recording their weight, neck size, length, and sex. They hoped to find a way to estimate weight from the other, more easily determined quantities.

28. **Schools** The State Education Department requires local school districts to keep these records on all students: age, race or ethnicity, days absent, current grade level, standardized test scores in reading and mathematics, and any disabilities or special educational needs.

29. **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, the number of calories, and the serving size in ounces. The data might be used to assess the nutritional value of the different sandwiches.

30. **Age and party** The Gallup Poll conducted a representative telephone survey of 1180 American voters during the first quarter of 2007. Among the reported results were the voter's region (Northeast, South, etc.), age, party affiliation, and whether or not the person had voted in the 2006 midterm congressional election.

31. **Babies** Medical researchers at a large city hospital investigating the impact of prenatal care on newborn health collected data from 882 births during 1998–2000. They kept track of the mother's age, the number of weeks the pregnancy lasted, the type of birth (cesarean, induced, natural), the level of prenatal care the mother had (none, minimal, adequate), the birth weight and sex of the baby, and whether the baby exhibited health problems (none, minor, major).

32. **Flowers** In a study appearing in the journal *Science*, a research team reports that plants in southern England are flowering earlier in the spring. Records of the first flowering dates for 385 species over a period of 47 years show that flowering has advanced an average of 15 days per decade, an indication of climate warming, according to the authors.

33. **Herbal medicine** Scientists at a major pharmaceutical firm conducted an experiment to study the effectiveness of an herbal compound to treat the common cold. They exposed each patient to a cold virus, then gave them either the herbal compound or a sugar solution known to have no effect on colds. Several days

Year	Winner	Jockey	Trainer	Owner	Time
1875	Aristides	O. Lewis	A. Williams	H. P. McGrath	2:37.75
1876	Vagrant	R. Swim	J. Williams	William Astor	2:38.25
1877	Baden Baden	W. Walker	E. Brown	Daniel Swigert	2:38
1878	Day Star	J. Carter	L. Paul	T. J. Nichols	2:37.25
...					
2010	Super Saver	C. Borel	T. Pletcher	WinStar Farm	2:04.04
2011	Animal Kingdom	J. Velazquez	H. G. Motion	Team Valor	2:02.04
2012	I'll Have Another	M. Gutierrez	D. O'Neill	Reddam Racing	2:01.83
2013	Orb	J. Rosario	S. McGaughey	Stuart Janney & Phipps Stable	2:02.89
2014	California Chrome	Victor Espinoza	Art Sherman	California Chrome, LLC	2:03.66
2015	American Pharoah	Victor Espinoza	Bob Baffert	Zayat Stables, LLC	2:03.03
2016	Nyquist	M. Gutierrez	Doug F. O'Neill	Reddam Racing LLC	2:01.31

Source: Excerpt from HorseHats.com. Published by Thoroughbred Promotions.

later they assessed each patient's condition, using a cold severity scale ranging from 0 to 5. They found no evidence of benefits of the compound.

34. Vineyards Business analysts hoping to provide information helpful to American grape growers compiled these data about vineyards: size (acres), number of years in existence, state, varieties of grapes grown, average case price, gross sales, and percent profit.

35. Streams In performing research for an ecology class, students at a college in upstate New York collect data on streams each year. They record a number of biological, chemical, and physical variables, including the stream name, the substrate of the stream (limestone, shale, or mixed), the acidity of the water (pH), the temperature (°C), and the BCI (a numerical measure of biological diversity).

36. Fuel economy The Environmental Protection Agency (EPA) tracks fuel economy of automobiles based on information from the manufacturers (Ford, Toyota, etc.). Among the data the agency collects are the manufacturer, vehicle type (car, SUV, etc.), weight, horsepower, and gas mileage (mpg) for city and highway driving.

37. Refrigerators In 2013, *Consumer Reports* published an article evaluating refrigerators. It listed 353 models, giving the brand, cost, size (cu ft), type (such as top freezer), estimated annual energy cost, an overall rating (good, excellent, etc.), and the repair history for that brand (percentage requiring repairs over the past 5 years).

38. Walking in circles People who get lost in the desert, mountains, or woods often seem to wander in circles rather than walk in straight lines. To see whether people naturally walk in circles in the absence of visual clues, researcher Andrea Axtell tested 32 people on a football field. One at a time, they stood at the center of one goal line, were blindfolded, and then tried to walk to the other goal line. She recorded each individual's sex, height, handedness, the number of yards each was able to walk before going out of bounds, and whether each wandered off course to the left or the right. No one made it all the way to the far end of the field without crossing one of the sidelines. (Source: *STATS* No. 39, Winter 2004)

39. Kentucky Derby 2016 The Kentucky Derby is a horse race that has been run every year since 1875 at Churchill Downs in 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 3-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.) Above are the data for the first four and seven recent races.

40. Indy 500 2016 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. In 2013, the winner, Tony Kanaan, averaged over 187 mph, beating the previous record by over 17 mph!

Here are the data for the first five races and five recent Indianapolis 500 races.

Year	Driver	Time (hr:min:sec)	Speed (mph)
1911	Ray Harroun	6:42:08	74.602
1912	Joe Dawson	6:21:06	78.719
1913	Jules Goux	6:35:05	75.933
1914	René Thomas	6:03:45	82.474
1915	Ralph DePalma	5:33:55.51	89.840
...			
2012	Dario Franchitti	2:58:51.2532	167.734
2013	Tony Kanaan	2:40:03.4181	187.433
2014	Ryan Hunter-Reay	2:40:48.2305	186.563
2015	Juan Pablo Montoya	3:05:56.5286	161.341
2016	Alexander Rossi	3:00:02.0872	166.634

41. Kentucky Derby 2016 on the computer Load the Kentucky Derby 2016 data into your preferred statistics package and answer the following questions;

a) What was the name of the winning horse in 1880?
b) When did the length of the race change?
c) What was the winning time in 1974?
d) Only one horse has run the Derby in less than 2 minutes. Which horse and in what year?

42. Indy 500 2016 on the computer Load the Indy 500 2016 data into your preferred statistics package and answer the following questions:

a) What was the average speed of the winner in 1920?
b) How many times did Bill Vukovich win the race in the 1950s?
c) How many races took place during the 1940s?

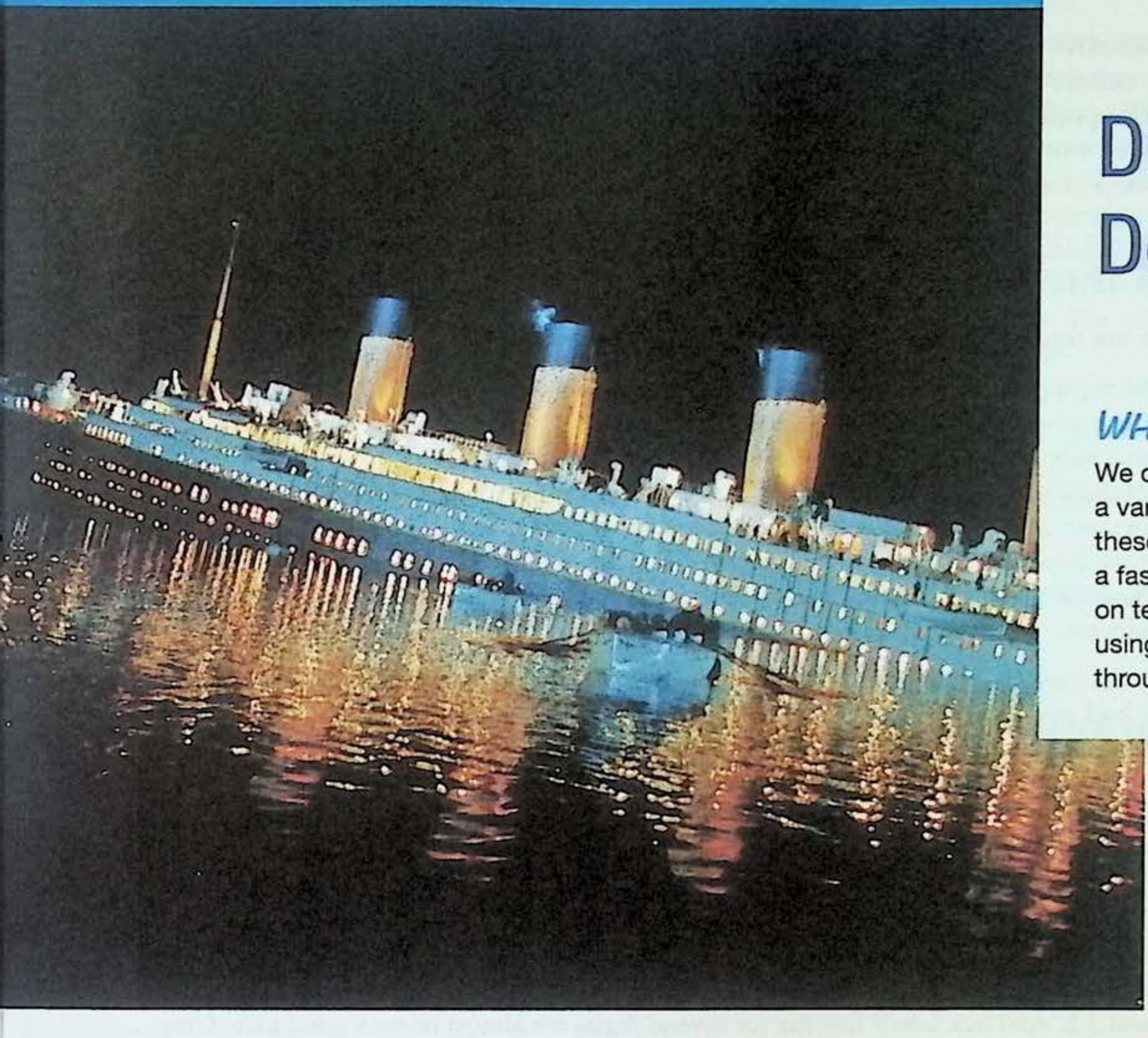
JUST CHECKING

Answers

1. *Who*—Tour de France races; *What*—year, winner, country of origin, age, team, total time, average speed, stages, total distance ridden, starting riders, finishing riders; *How*—official statistics at race; *Where*—France (for the most part); *When*—1903 to 2016; *Why*—not specified (To see progress in speeds of cycling racing?)

2.

Variable	Type	Units
Year	Quantitative or Identifier	Years
Winner	Categorical	
Country of Origin	Categorical	
Age	Quantitative	Years
Team	Categorical	
Total Time	Quantitative	Hours/minutes/ seconds
Average Speed	Quantitative	Kilometers per hour
Stages	Quantitative	Counts (stages)
Total Distance	Quantitative	Kilometers
Starting Riders	Quantitative	Counts (riders)
Finishing Riders	Quantitative	Counts (riders)



2

Displaying and Describing Data

WHERE ARE WE GOING?

We can summarize and describe data values in a variety of ways. You'll probably recognize these displays and summaries. This chapter is a fast review of these concepts so we all agree on terms, notation, and methods. We'll be using these displays and descriptions throughout the rest of the book.

What happened on the *Titanic* at 11:40 on the night of April 14, 1912, is well known. Frederick Fleet's cry of "Iceberg, right ahead" and the three accompanying pulls of the crow's nest bell signaled the beginning of a nightmare that has become legend. By 2:15 AM, the *Titanic*, thought by many to be unsinkable, had sunk. Only 712 of the 2208 people on board survived. The others (nearly 1500) met their icy fate in the cold waters of the North Atlantic.

Table 2.1 shows some data about the passengers and crew aboard the *Titanic*. Each case (row) of the data table represents a person on board the ship. The variables are the person's *Name, Survival* status (Dead or Alive), *Age* (in years), *Age Category* (Adult or Child), *Sex* (Male or Female), *Price* Paid (in British pounds, **£**), and ticket *Class* (First, Second, Third, or Crew). Some of these, such as *Age* and *Price*, record numbers. These are called

Table 2.1
Part of a data table showing seven variables for 11 people aboard the *Titanic*.

Name	Survived	Age	Adult/Child	Sex	Price (£)	Class
ABBING, Mr Anthony	Dead	42	Adult	Male	7.55	3
ABBOTT, Mr Ernest Owen	Dead	21	Adult	Male	0	Crew
ABBOTT, Mr Eugene Joseph	Dead	14	Child	Male	20.25	3
ABBOTT, Mr Rossmore Edward	Dead	16	Adult	Male	20.25	3
ABBOTT, Mrs Rhoda Mary "Rosa"	Alive	39	Adult	Female	20.25	3
ABELSETH, Miss Karen Marie	Alive	16	Adult	Female	7.65	3
ABELSETH, Mr Olaus Jörgensen	Alive	25	Adult	Male	7.65	3
ABELSON, Mr Samuel	Dead	30	Adult	Male	24	2
ABELSON, Mrs Hannah	Alive	28	Adult	Female	24	2
ABRAHAMSSON, Mr Abraham August Johannes	Alive	20	Adult	Male	7.93	3
ABRAHIM, Mrs Mary Sophie Halaut	Alive	18	Adult	Female	7.23	3

quantitative variables. Others, like *Survival* and *Class*, place each case in a single category, and are called **categorical** variables. (Data in **Titanic**)

The problem with a data table like this—and in fact with all data tables—is that you can't *see* what's going on. And seeing is just what we want to do. We need ways to show the data so that we can see patterns, relationships, trends, and exceptions.

WHO	People on the *Titanic*
WHAT	Name, survival status, age, adult/child, sex, price paid, ticket class
WHEN	April 14, 1912
WHERE	North Atlantic
HOW	www.encyclopedia-titanica.org
WHY	Historical interest

The Three Rules of Data Analysis

There are three things you should always do first with data:

1. **Make a picture.** A display of your data will reveal things you're not likely to see in a table of numbers and will help you *Think* clearly about the patterns and relationships that may be hiding in your data.
2. **Make a picture.** A well-designed display will *Show* the important features and patterns in your data. It could also show you things you did not expect to see: extraordinary (possibly wrong) data values or unexpected patterns.
3. **Make a picture.** The best way to *Tell* others about your data is with a well-chosen picture.

These are the three rules of data analysis. There are pictures of data throughout the book, and new kinds keep showing up. These days, technology makes drawing pictures of data easy, so there is no reason not to follow the three rules.

We make graphs for two primary reasons: to understand more about data and to show others what we have learned and want them to understand. The first reason calls for simple graphs with little adornment; the second often uses visually appealing additions to draw the viewer's attention. Regardless of their function, graphs should be easy to read and understand and should represent the facts of the data honestly. Axes should be clearly labeled with the names of the variables they display. The intervals set off by "tick marks" should occur at values easy to think about: 5, 10, 15, and 20 are simpler marks than, say, 1.7, 2.3, 2.9, and 3.5. And tick labels that run for several digits are almost never a good idea. Graphs should have a "key" that identifies colors and symbols if those are meaningful in the graph. And all graphs should carry a title or caption that says what the graph displays and suggests what about it is salient or important.

The Area Principle

A bad picture can distort our understanding rather than help it. What impression do you get from Figure 2.1 about who was aboard the ship?

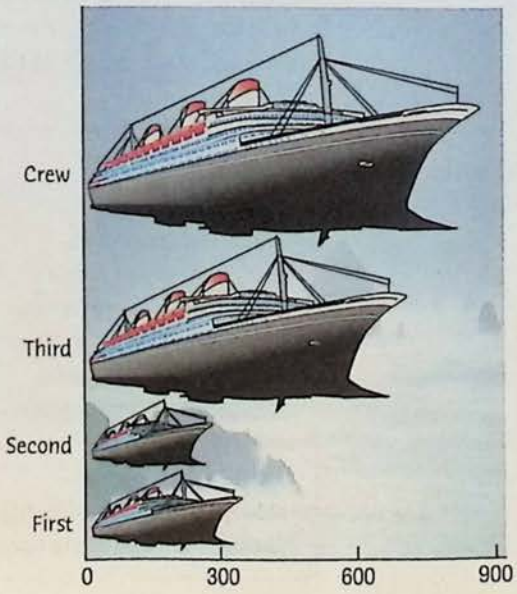


Figure 2.1
How many people were in each class on the *Titanic*? From this display, it looks as though the service must have been great, since most aboard were crew members. Although the length of each ship here corresponds to the correct number, the impression is all wrong. In fact, only about 40% were crew.

The *Titanic* was certainly a luxurious ship, especially for those in first class, but Figure 2.1 gives the mistaken impression that most of the people on the *Titanic* were crew members, with a few passengers along for the ride. What's wrong? The lengths of the ships *do* match the number of people in each ticket class category. However, our eyes tend to be more impressed by the *area* than by other aspects of each ship image. So, even though the *length* of each ship matches up with one of the totals, it's the associated *area* in the image that we notice. There were about 3 times as many crew as second-class passengers, and the ship depicting the number of crew members is about 3 times longer than the ship depicting second-class passengers. The problem is that it occupies about 9 times the area. That just isn't a correct impression.

The best data displays observe a fundamental principle of graphing data called the **area principle**. The area principle says that the area occupied by a part of the graph should correspond to the magnitude of the value it represents. Violations of the area principle are a common way to lie (or, since most mistakes are unintentional, we should say err) with statistics.

2.1 Summarizing and Displaying a Categorical Variable

Frequency Tables

Categorical variables are easy to summarize in a **frequency table** that lists the number of cases in each category along with its name.

For ticket *Class*, the categories are First, Second, Third, and Crew:

Class	Count
First	324
Second	285
Third	710
Crew	889

Table 2.2

A frequency table of the *Titanic* passengers.

Class	Percentage (%)
First	14.67
Second	12.91
Third	32.16
Crew	40.26

Table 2.3

A relative frequency table for the same data.

A **relative frequency table** displays *percentages* (or *proportions*) rather than the counts in each category. Both types of tables show the **distribution** of a categorical variable because they name the possible categories and tell how frequently each occurs. (The percentages should total 100%, although the sum may be a bit too high or low if the individual category percentages have been rounded.)

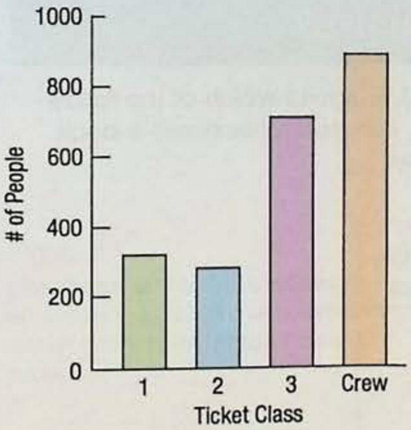


Figure 2.2

People on the Titanic *by Ticket Class.* With the area principle satisfied, we can see the true distribution more clearly.

Bar Charts

Although not as visually entertaining as the ships in Figure 2.1, the **bar chart** in Figure 2.2 gives an *accurate* visual impression of the distribution because it obeys the area principle. Now it's easy to see that the majority of people on board were *not* crew. We can also see that there were about 3 times as many crew members as second-class passengers. And there were more than twice as many third-class passengers as either first- or second-class passengers—something you may have missed in the frequency table. Bar charts make these kinds of comparisons easy and natural.

EXAMPLE 2.1

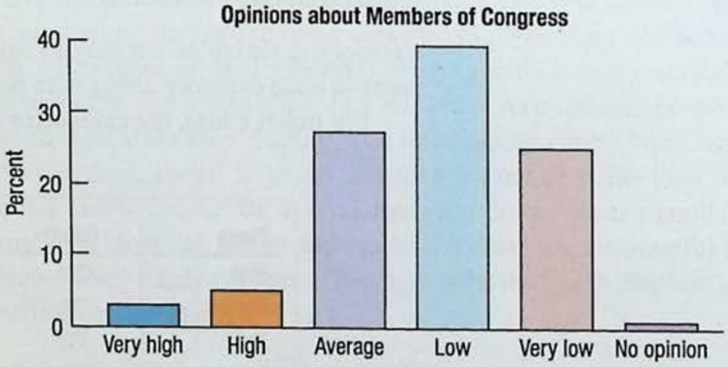
What Do You Think of Congress?

In December 2015, the Gallup survey asked 824 people how they viewed a variety of professions. Specifically they asked, "How would you rate the honesty and ethical standards of people in these different fields?" For Members of Congress, the results were

Rating	Percentage (%)
Very high	3
High	5
Average	27
Low	39
Very low	25
No opinion	1

QUESTION: What kind of table is this? What would be an appropriate display?

ANSWER: This is a relative frequency table because the numbers displayed are percentages, not counts. A bar chart would be appropriate:



EXAMPLE 2.2

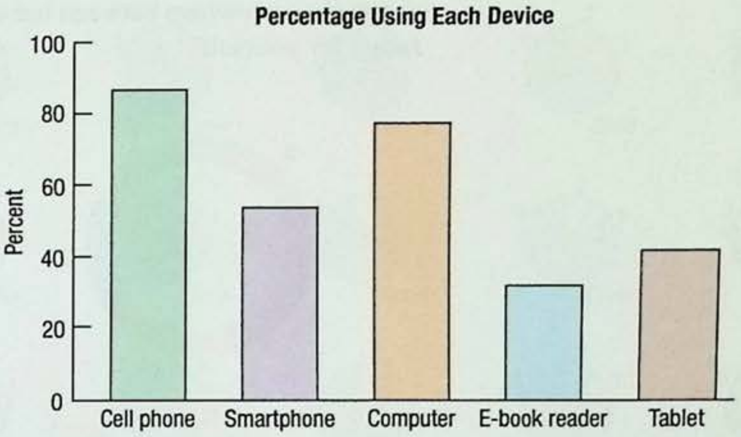
Which Gadgets Do You Use?

In 2014, the Pew Research Organization asked 1005 U.S. adults which of the following electronic items they use: cell phone, smartphone, computer, handheld e-book reader (e.g., Kindle or Nook), or tablet. The results were

Device	Percentage (%) using the device
Cell phone	86.8
Smartphone	54.0
Computer	77.5
E-book reader	32.2
Tablet	41.9

QUESTION: Is this a frequency table, a relative frequency table, or neither? How could you display these data?

ANSWER: This is not a frequency table because the numbers displayed are not counts. Although the numbers are percentages, they do not sum to 100%. A person can use more than one device, so this is not a relative frequency table either. A bar chart might still be appropriate, but the numbers do not sum to 100%.

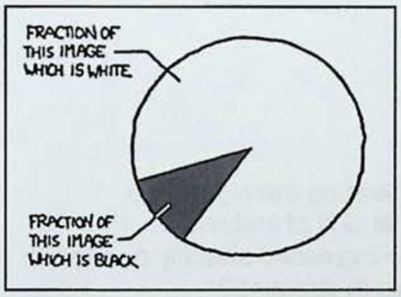


Pie Charts

Pie charts display all the cases as a circle whose slices have areas proportional to each category's fraction of the whole.

Pie charts give a quick impression of the distribution. Because we're used to cutting up pies into 2, 4, or 8 pieces, pie charts are particularly good for seeing relative frequencies near 1/2, 1/4, or 1/8.

Bar charts are almost always better than pie charts for comparing the relative frequencies of categories. Pie charts are widely understood and colorful, and they often appear in reports, but Figure 2.3 shows why statisticians prefer bar charts.


FRACTION OF THIS IMAGE WHICH IS WHITE
FRACTION OF THIS IMAGE WHICH IS BLACK

© 2013 Randall Munroe. Reprinted with permission. All rights reserved.

Figure 2.3
Pie charts may be attractive, but it can be hard to see patterns in them. Can you discern the differences in distributions depicted by these pie charts?

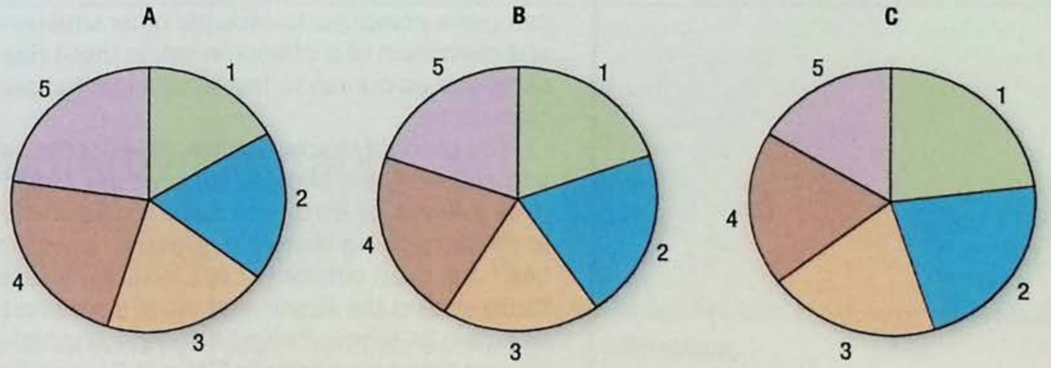
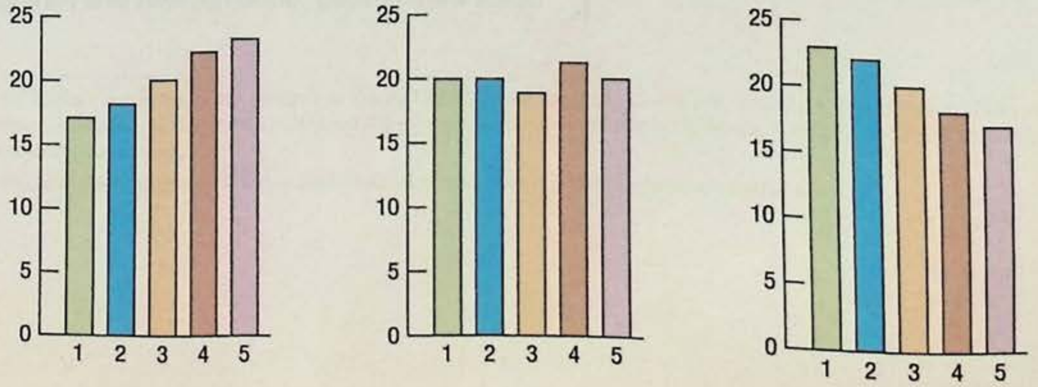


Figure 2.4
Bar charts of the same values as shown in Figure 2.3 make it much easier to compare frequencies in groups.



Ring Charts

A ring (or donut) chart is a modified form of pie chart that displays only the "crust" of the pie—a ring that is partitioned into regions proportional in area to each value. You can think of the ring as the bars of a bar chart stuck end to end and wrapped around the circle. Ring charts are somewhere between bar charts and pie charts. They may be easier to read (or not). Judge for yourself:

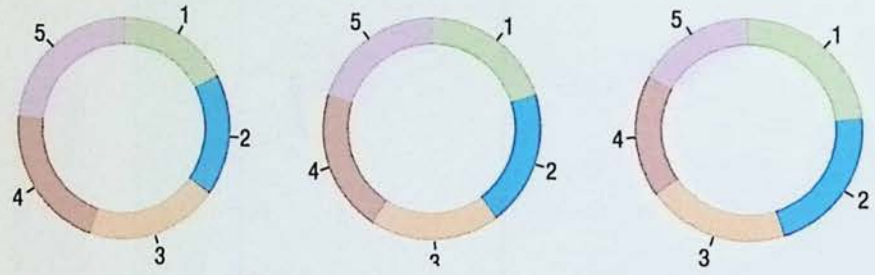


Figure 2.5

Ring charts compromise between pie and bar charts. These ring charts show the same values as the pie charts in Figure 2.3. Do you find it easier to see the patterns?

RANDOM MATTERS

Is it random, or is something systematic going on? Separating the *signal* (the systematic) from the *noise* (the random) is a fundamental skill of statistics.

A geoscientist notices that global temperatures have increased steadily during the past 50 years. Could the pattern be random, or is the earth warming?

An analyst notices that the stock market seems to go up more often on Tuesday afternoons when it rains in Chicago. Is that something she should bank on?

One of the challenges to answering questions like these is that we have only one earth and one stock market history. What if we had two? Or many? Sometimes we can use a computer to *simulate* other situations, to pretend that we have more than one realization of a phenomenon. In these *Random Matters* sections, we'll use the computer as our lab to test what might happen if we could repeat our data collection many times.

You probably know that the "rules of the sea" were enforced on the *Titanic*—women and children were allowed to board the *Titanic* lifeboats before the men. Did ticket class (first, second, or third) also make a difference? Suppose the 712 survivors were chosen at random, giving everyone an equal chance to get into a lifeboat. Would the distribution have been different? Let's look. We selected 712 people at random from the list of those aboard the *Titanic* and made a pie chart of ticket class. We repeated the random selection 24 times, making a new pie chart of each selected group of 712 passengers. Among these pie charts in Figure 2.6 we've "hidden" the actual distribution of survivors. Can you pick out the real distribution? If so, then that might convince you that the lifeboats weren't filled randomly, with everyone getting an equal chance.

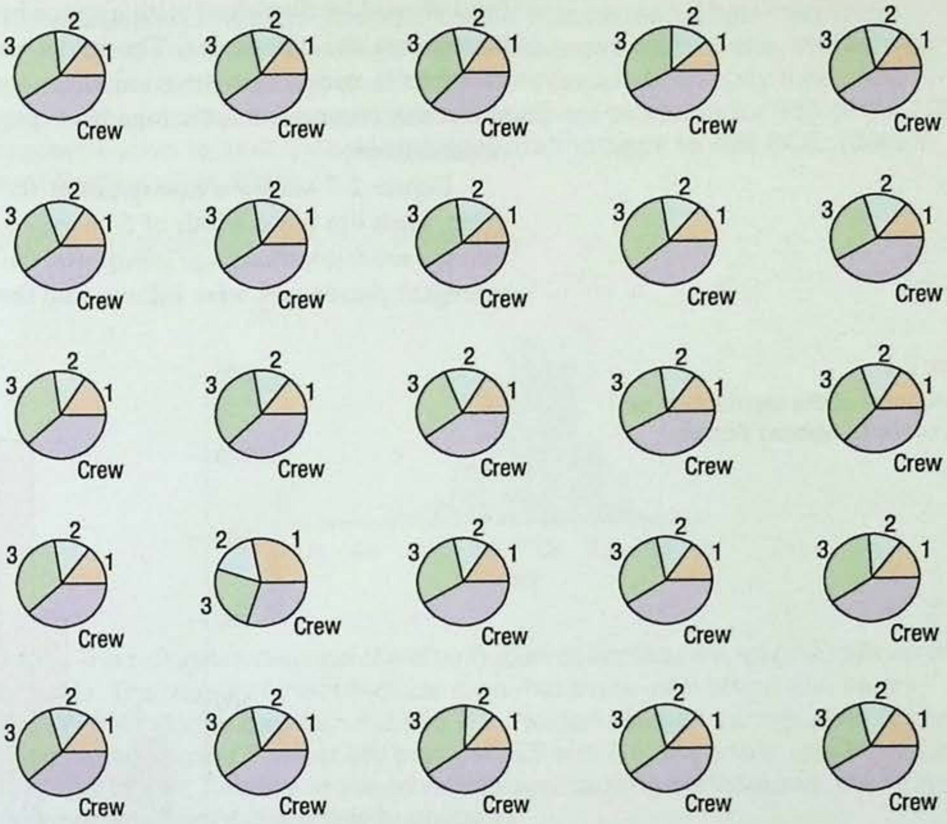


Figure 2.6

The distribution of ticket class in 24 simulated lifeboats and the actual distribution of survivors. Can you find the real one? If so, this suggests that people didn't all have an equal chance to survive.[1]

In this example, the difference is pretty obvious. There were more survivors from first and second class and fewer from third class than there would have been were everyone given an equal chance. In other situations, the differences may not be as obvious, so we'll need to develop more sophisticated tools to help distinguish signals from noise.

2.2 Displaying a Quantitative Variable

Histograms

How can we make a bar chart for a quantitative variable? We can't, because quantitative variables don't have categories. Instead, we make a **histogram**.

Histograms and bar charts both use bars, but they are fundamentally different. The bars of a bar chart display the count for each category, so they could be arranged in any order[2]

[1] Wait. Didn't we just say we prefer bar charts? Well, sometimes pie charts are actually a good choice. They are compact, colorful, and—most important—they satisfy the area principle. A figure with 25 bar charts would look much more confusing.

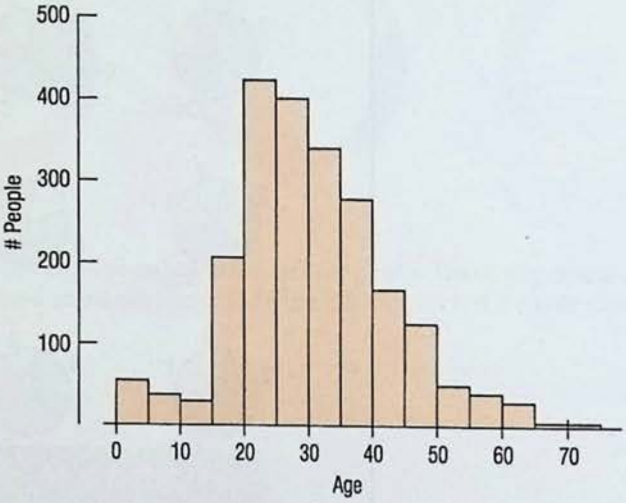
[2] Many statistics programs choose alphabetical order, which is rarely the most useful one.

(and should be displayed with a space between them). The horizontal axis of a bar chart just names the categories. The horizontal axis of a histogram shows the values of the variable in order. A histogram slices up that axis into equal-width bins, and the bars show the counts for each bin. Now **gaps** are meaningful; they show regions with no observations.

Figure 2.7 shows a histogram of the ages of those aboard the *Titanic*. In this histogram, each bin has a width of 5 years, so, for example, the height of the tallest bar shows that the most populous age group was the 20- to 24-year-olds, with over 400 people.[3] The youngest passengers were infants, and the oldest was more than 70 years old.

Figure 2.7

A histogram of the distribution of ages of those aboard *Titanic*.



The fact that there are fewer and fewer people in the 5-year bins from age 25 on probably doesn't surprise you either. After all, there are increasingly fewer people of advancing age in the general population as well, and there were no very elderly people on board the *Titanic*. But the bins on the left are a little strange. It looks like there were more infants and toddlers (0–5 years old) than there were preteens.

Does this distribution look plausible? You may not have guessed that fact about the infants and preteens, but it doesn't seem out of the question. It is often a good idea to imagine what the distribution might look like before you make the display. That way you'll be less likely to be fooled by errors in the data or when you accidentally graph the wrong variable.[4]

EXAMPLE 2.3

Earthquakes and Tsunamis

In 2011, the most powerful earthquake ever recorded in Japan created a wall of water that devastated the northeast coast of Japan and left nearly 25,000 people dead or missing. The 2011 tsunami in Japan was caused by a 9.0 magnitude earthquake. It was particularly noted for the damage it caused to the Fukushima Daiichi nuclear power plant, causing a core meltdown and international concern. As disastrous as it was, the Japan tsunami was not nearly as deadly as the 2004 tsunami on the west coast of Sumatra that killed an estimated 227,899 people, making it the

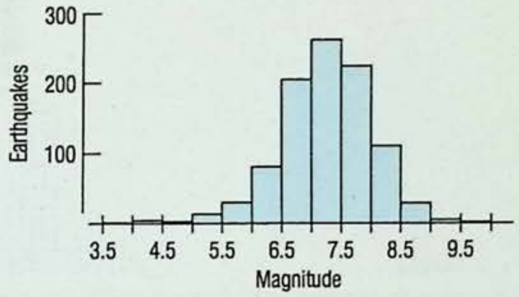
[3] The histogram bar appears to go from 20 to 25, but most statistics programs include values at the lower limit in the bar and put values at the upper limit in the next bin.

[4] You'll notice that we didn't say *if* you graph the wrong variable, but rather *when*. Everyone makes mistakes, and you'll make your share. But if you always think about what your graph or analysis says about the world and judge whether that is reasonable, you can catch many errors before they get away.

most lethal tsunami on record. The earthquake that caused it had magnitude 9.1—more than 25% more powerful than the Japanese earthquake. Were these earthquakes truly extraordinary, or did they just happen at unlucky times and places? The magnitudes (measured or estimated) are available for 968 of the 1087 earthquakes known to have caused tsunamis, dating back to 426 BCE. (Data in **Tsunami Earthquakes 2016**).

WHO	1087 earthquakes known to have caused tsunamis for which we have data or good estimates
WHAT	Magnitude (Richter scale), depth (m), date, location, and other variables
WHEN	From 426 BCE to the present
WHERE	All over the earth

QUESTION: What can we learn from these data?



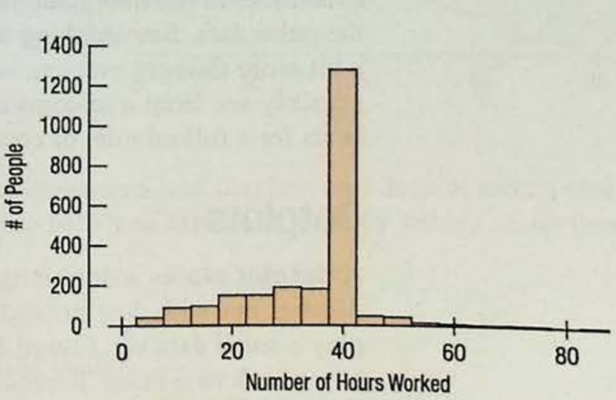
ANSWER: The histogram displays the distribution of earthquake magnitudes on the Richter scale. The height of the tallest bar says that there were about 250 earthquakes with magnitudes between 7.0 and 7.5. We can see that earthquakes typically have magnitudes around 7. Most are between 5.5 and 8.5, but one is less than 4 and a few are 9 or bigger. Relative to the other tsunami-causing earthquakes, the Sumatra and Japan events were extraordinarily powerful.

EXAMPLE 2.4

How Much Do Americans Work?

The Bureau of Labor Statistics (BLS) collects data on many aspects of the U.S. economy. One of the surveys it conducts, the American Time Use Survey (ATUS), asks roughly 11,000 people a year a variety of questions about how they spend their time. For those who are employed, it asks how many hours a week they work. Here is a histogram of the 2270 responses in 2014.

QUESTION: What does the histogram say about how many hours U.S. workers typically work.

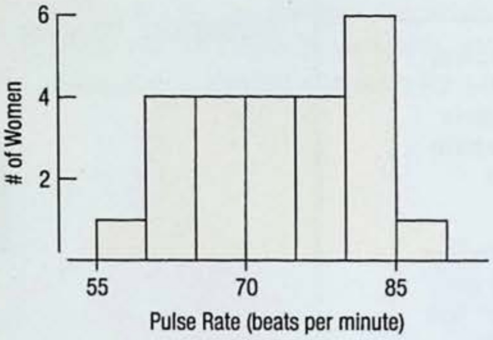


ANSWER: It looks like the vast majority of people (more than 1200 in this study) work right around 40 hours a week. There are some who work less, and a very few who work more.

Stem-and-Leaf Displays

Histograms provide an easy-to-understand summary of the distribution of a quantitative variable, but they don't show the data values themselves. For example, here's a histogram of the pulse rates of 24 women at a health clinic:

Figure 2.8
A histogram of the pulse rates of 24 women at a health clinic.



Here's a stem-and-leaf display of the same data:

```
5 | 6
6 | 0444
6 | 8888
7 | 2222
7 | 6666
8 | 000044
8 | 8
  Pulse Rate
(5|6 means 56 beats/min)
```

STEM-AND-LEAF OR STEMPLOT?

The stem-and-leaf display was devised by John W. Tukey, one of the greatest statisticians of the 20th century. It is called a "stemplot" in some texts and computer programs.

A **stem-and-leaf display** is like a histogram, but it shows the individual values. It's also easier to make by hand. Turn the stem-and-leaf on its side (or turn your head to the right) and squint at it. It should look roughly like the histogram of the same data. Does it?[5]

The first line of the display, which says 5 | 6, stands for a pulse of 56 beats per minute (bpm). We've taken the tens place of the number and made that the "stem." Then we sliced off the ones place and made it a "leaf." The next line down is 6 | 0444, which shows one pulse rate of 60 and three of 64 bpm.

Stem-and-leaf displays are especially useful when you make them by hand for batches of fewer than a few hundred data values. They are a quick way to display—and even to record—numbers. Because the leaves show the individual values, we can sometimes see even more in the data than the distribution's shape. Take another look at all the leaves of the pulse data. See anything unusual? At a glance you can see that they are all even. With a bit more thought you can see that they are all multiples of 4—something you couldn't possibly see from a histogram. How do you think the nurse took these pulses? Counting beats for a full minute, or counting for only 15 seconds and multiplying by 4?

Dotplots

A **dotplot** places a dot along an axis for each case in the data. It's like a stem-and-leaf display, but with dots instead of digits for all the leaves. Dotplots are a great way to display a small data set. Figure 2.9 shows a dotplot of the time (in seconds) that the winning horse took to win the Kentucky Derby in each race between the first Derby in 1875 and the 2015 Derby.

[5] You could make the stem-and-leaf display with the higher values on the top. Putting the lower values at the top matches the histogram; putting the higher values at the top matches the way a vertical axis works in other displays such as dotplots (as we'll see presently).

Dotplots display basic facts about the distribution. We can find the slowest and fastest races by finding the times for the topmost and bottommost dots. It's clear that there are two clusters of points, one just below 160 seconds and the other at about 122 seconds. Something strange happened to the Derby times. Once we know to look for it, we can find out that in 1896 the distance of the Derby race was changed from 1.5 miles to the current 1.25 miles. That explains the two clusters of winning times.

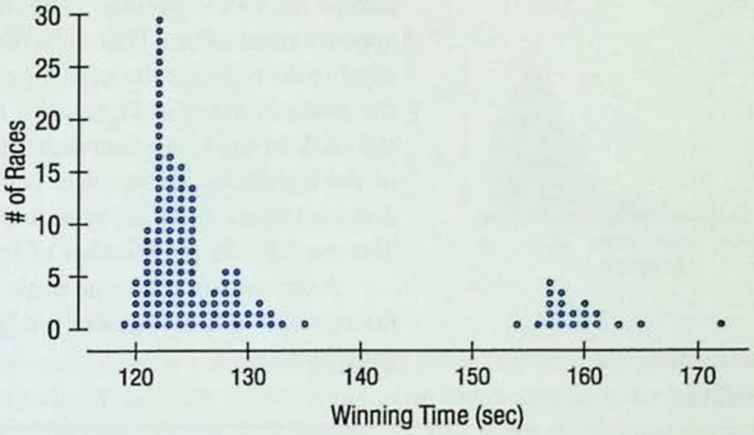


Figure 2.9
A dotplot of Kentucky Derby winning times plots each race as its own dot. We can see two distinct groups corresponding to the two different race distances.

WHO	Runnings of the Kentucky Derby
WHAT	Winning time
WHEN	1875–2016
WHERE	Churchill Downs

*Density Plots

The size of the bins in a histogram can influence its look and our interpretation of the distribution. There is no correct bin size, although recommendations to use between 5 and 20 bins are common. **Density plots** smooth the bins in a histogram to reduce the effect of this choice. How much the bin heights are smoothed is still a choice that affects the shape, but the change in shape is less severe than in a histogram. Here's a density plot of the *Ages* of those on the *Titanic*. Compare it to Figure 2.7.

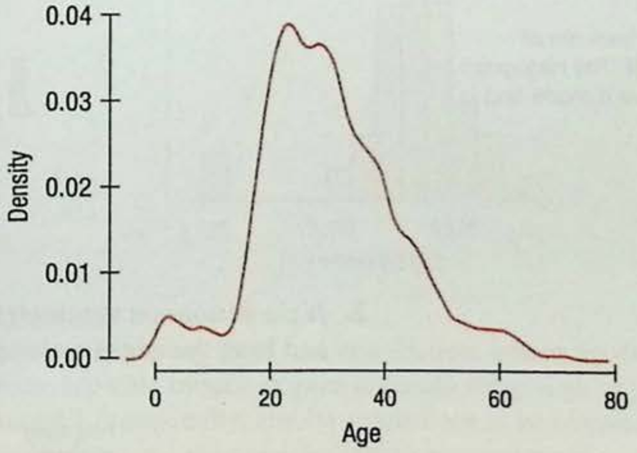


Figure 2.10
A density plot of the *Ages* of those aboard the *Titanic*. We can see, as we did in the histogram, that the most populous age is near 20 and that there are more infants and toddlers than preteens. The density plot does not provide hard cut-offs to the bins, but smooths the distribution over the bins.

Every histogram, stem-and-leaf display, and dotplot tells a story, but you need to develop a vocabulary to help you explain it. Start by talking about three things: its *shape*, *center*, and *spread*.

Think Before You Draw

Before making a pie chart or a bar chart, you should check that you have categorical data. Before making a stem-and-leaf display, a histogram, or a dotplot, you should make sure you are working with quantitative data. Although a bar chart and a histogram may look similar, they're not the same display. You can't display categorical data in a histogram nor quantitative data in a bar chart.

2.3 Shape

We summarize the **shape** of a distribution in terms of three attributes: how many *modes* it has, whether it is *symmetric* or *skewed*, and whether it has any extraordinary cases or *outliers*.

1. *Does the histogram have a single, central hump or several separated humps?* These humps are called **modes**[6]. The mode is sometimes defined as the single value that appears most often. That definition is fine for categorical variables because all we need to do is count the number of cases for each category. For quantitative variables, the mode is more ambiguous. It makes more sense to use the term "mode" to refer to the peak of the histogram rather than as a single summary value. The important feature of the Kentucky Derby races is that there are two distinct modes, representing the two different versions of the race and warning us to consider those two versions separately. The earthquake magnitudes of Example 2.3 have a single mode at just about 7.

   A histogram with one peak, such as the ages (Figure 2.7), is dubbed **unimodal**; histograms with two peaks are **bimodal**, and those with three or more are called **multimodal**.[7]

> ### PIE À LA MODE?
>
> You've heard of pie à la mode. Is there a connection between pie and the mode of a distribution? Actually, there is! The mode of a distribution is a *popular* value near which a lot of the data values gather. And "à la mode" means "in style"—*not* "with ice cream." That just happened to be a *popular* way to have pie in Paris around 1900.

A histogram that doesn't appear to have any mode and in which all the bars are approximately the same height is called **uniform**.

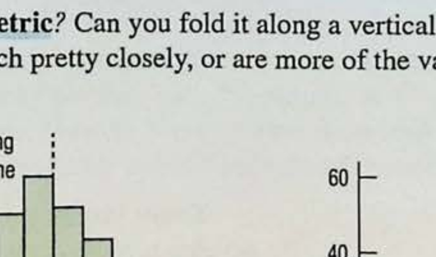


Figure 2.11
In this histogram, the bars are all about the same height. The histogram doesn't appear to have a mode and is called uniform.

2. *Is the histogram* **symmetric**? Can you fold it along a vertical line through the middle and have the edges match pretty closely, or are more of the values on one side?

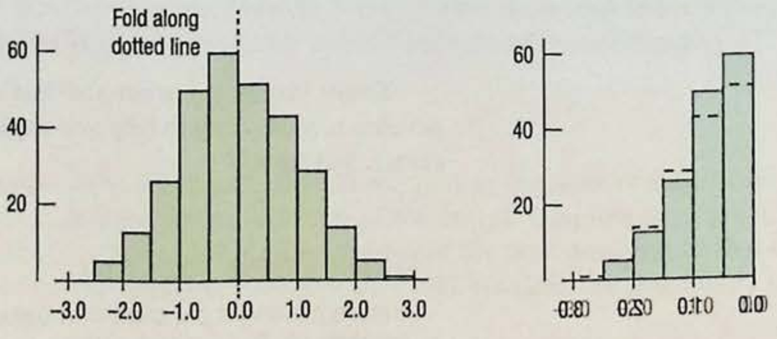


Figure 2.12
A symmetric histogram can fold in the middle so that the two sides almost match.

[6] Well, technically, it's the value on the horizontal axis of the histogram that is the mode, but anyone asked to point to the mode would point to the hump.

[7] Apparently, statisticians don't like to count past two.

We generally prefer to work with symmetric distributions because they are easier to summarize, model, and discuss. For example, it is pretty clear where the center of the distribution is located.

The (usually) thinner ends of a distribution are called the **tails**. If one tail stretches out farther than the other, the histogram is said to be **skewed** to the side of the longer tail.

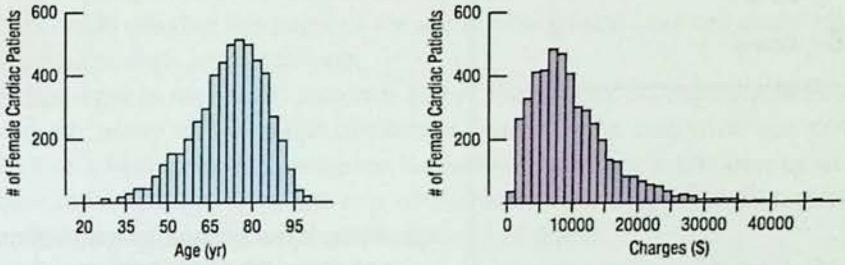


Figure 2.13
Two skewed histograms showing data on two variables for all female heart attack patients in New York State in one year. The blue one (age in years) is skewed to the left. The purple one (charges in $) is skewed to the right.

3. *Do any unusual features stick out?* Often such features say something interesting or exciting about the data. Always mention any stragglers, or **outliers**, that stand away from the body of the distribution. If you were collecting data on nose lengths and Pinocchio was in the group, he'd be an outlier, and you'd certainly want to mention it.

    Outliers can affect almost every method we discuss in this course. So we'll always be on the lookout for them. An outlier can be the most informative part of your data. Or it might just be an error. Never throw data away without comment. Treat outliers specially and discuss them when you tell about your data. Or find the error and fix it if you can. Be sure to look for outliers. Always.

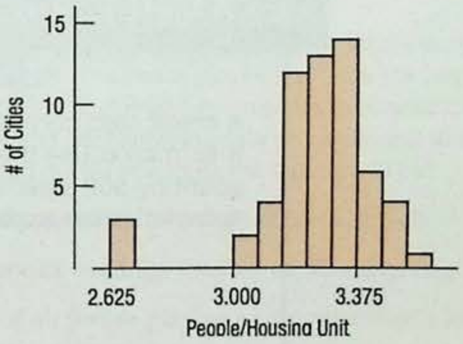


Figure 2.14
A histogram of the number of people per housing unit in a sample of cities. The three cities in the leftmost bar are outliers. We wonder why they are different.

You should also notice any gaps in a distribution, where no data appear at all. Gaps may indicate separate modes or even separate subgroups in your data that you may wish to consider individually. But be careful not to be overzealous when looking for gaps. For small data sets, they may be due to happenstance.

EXAMPLE 2.5 Consumer Price Index

The Consumer Price Index (CPI) summarizes the cost of a representative market basket of goods that includes groceries, restaurants, transportation, utilities, and medical care. Global companies often use the CPI to determine living allowances and salaries for employees. Inflation is often measured by how much the CPI changes from year to year. Relative CPIs can be found for different cities. We have data giving CPI components relative to New York City. For New York City, each index is 100(%). (Data in **CPI Worldwide 2016**)

Pearson

Get the Most Out of MyStatLab®

MyStatLab is the leading online homework, tutorial, and assessment program designed to help you learn and understand statistics.

- ✓ Personalized and adaptive learning
- ✓ Interactive practice with immediate feedback
- ✓ Multimedia learning resources
- ✓ Complete eText
- ✓ Mobile-friendly design
- ✓ Full access to StatCrunch

MyStatLab is available for this textbook.
To learn more, visit www.mystatlab.com

X5PT-89IG-6XGS
Now for the good stuff!
< Scan me
Pearson
INTERCEPTAG

www.pearson.com

ISBN-13: 978-0-13-421022-3
ISBN-10: 0-13-421022-0

EAN
9 780134 210223
90000>