CHAPTER 2

Data

"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, Director of Data Mining and Personalization, Amazon.com

Many years ago, most stores in small towns knew their customers personally. If you walked into the hobby shop, the owner might tell you about a new bridge that had come in for your Lionel train set. The tailor knew your dad's size, and the hairdresser knew how your mom liked her hair. There are still some stores like that around today, but we're increasingly likely to shop at large stores, by phone, or on the Internet. Even so, when you phone an 800 number to buy new running shoes, customer service representatives may call you by your first name or ask about the socks you bought 6 weeks ago. Or the company may send an e-mail in October offering new head warmers for winter running. This company has millions of customers, and you called without identifying yourself. How did the sales rep know who you are, where you live, and what you had bought?

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

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 2006, the company's revenue reached $10.7 billion. Amazon has expanded into selling a wide selection of merchandise, from $400,000 necklaces[1] to yak cheese from Tibet to the largest book in the world.

Amazon is constantly monitoring and evolving its Web site to serve its customers better and maximize sales performance. To decide which changes to make to the site, the company experiments, collecting data and analyzing what works best. When you visit the Amazon Web site, you may encounter a different look or different suggestions and offers. Amazon statisticians want to know whether you'll follow the links offered, purchase the items suggested, or even spend a

[1] Please get credit card approval before purchasing online.

longer time browsing the site. As Ronny Kohavi, director of Data Mining and Personalization, said, "Data trumps intuition. Instead of using our intuition, we experiment on the live site and let our customers tell us what works for them."

But What *Are* Data?

THE W'S:
WHO
WHAT
and in what units
WHEN
WHERE
WHY
HOW

We bet you thought you knew this instinctively. Think about it for a minute. What exactly *do* we mean by "data"?

Do data have to be numbers? The amount of your last purchase in dollars is numerical data, but some data record names or other labels. The names in Amazon.com's database are data, but not numerical.

Sometimes, data can have values that look like numerical values but are just numerals serving as labels. This can be confusing. For example, the ASIN (Amazon Standard Item Number) of a book, like 0321570448, may have a numerical value, but it's really just another name for *Stats: Modeling the World.*

Data values, no matter what kind, are useless without their context. Newspaper journalists know that the lead paragraph of a good story should establish the "Five W's": *Who, What, When, Where*, and (if possible) *Why.* Often we add *How* to the list as well. Answering these questions can provide the **context** for data values. The answers to the first two questions are essential. If you can't answer *Who* and *What,* you don't have **data,** and you don't have any useful information.

Data Tables

Here are some data Amazon might collect:

B000001OAA	10.99	Chris G.	902	15783947	15.98	Kansas	Illinois	Boston
Canada	Samuel P.	Orange County	N	B000068ZVQ	Bad Blood	Nashville	Katherine H.	N
Mammals	10783489	Ohio	N	Chicago	12837593	11.99	Massachusetts	16.99
312	Monique D.	10675489	413	B00000I5Y6	440	B000002BK9	Let Go	Y

A S ***Activity:*** **What Is (Are) Data?** Do you really know what's data and what's just numbers?

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

Purchase Order	Name	Ship to State/Country	Price	Area Code	Previous CD Purchase	Gift?	ASIN	Artist
10675489	Katharine H.	Ohio	10.99	440	Nashville	N	B00000I5Y6	Kansas
10783489	Samuel P.	Illinois	16.99	312	Orange County	Y	B000002BK9	Boston
12837593	Chris G.	Massachusetts	15.98	413	Bad Blood	N	B000068ZVQ	Chicago
15783947	Monique D.	Canada	11.99	902	Let Go	N	B000001OAA	Mammals

Now we can see that these are four purchase records, relating to CD orders from Amazon. The column titles tell *What* has been recorded. The rows tell us *Who.* But be careful. Look at all the variables to see *Who* the variables are about. Even if people are involved, they may not be the *Who* of the data. For example, the *Who* here are the purchase orders (not the people who made the purchases).

A common place to find the *Who* of the table is the leftmost column. The other W's might have to come from the company's database administrator.[2]

Who

In general, the rows of a data table correspond to individual **cases** about *Who*m (or about which—if they're not people) we record some characteristics. These cases go by different names, depending on the situation. Individuals who answer a survey are referred to as *respondents*. People on whom we experiment are *subjects* or (in an attempt to acknowledge the importance of their role in the experiment) *participants*, but animals, plants, Web sites, and other inanimate subjects are often just called *experimental units*. In a database, rows are called *records*—in this example, purchase records. Perhaps the most generic term is **cases.** In the Amazon table, the cases are the individual CD orders.

Sometimes people just refer to data values as *observations*, without being clear about the *Who*. Be sure you know the *Who* of the data, or you may not know what the data say.

Often, the cases are a **sample** of cases selected from some larger **population** that we'd like to understand. Amazon certainly cares about its customers, but also wants to know how to attract all those other Internet users who may never 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.

A S ***Activity:* Consider the Context . . .** Can you tell who's *Who* and what's *What?* And *Why?* This activity offers real-world examples to help you practice identifying the context.

FOR EXAMPLE **Identifying the "Who"**

In March 2007, *Consumer Reports* published an evaluation of large-screen, high-definition television sets (HDTVs). The magazine purchased and tested 98 different models from a variety of manufacturers.

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

The magazine is interested in the performance of all HDTVs currently being offered for sale. It tested a sample of 98 sets, the "Who" for these data. Each HDTV set represents all similar sets offered by that manufacturer.

What and Why

The characteristics recorded about each individual are called **variables.** These are usually shown as the columns of a data table, and they should have a name that identifies *What* has been measured. Variables may seem simple, but to really understand your variables, you must *Think* about what you want to know.

Although area codes are numbers, do we use them that way? Is 610 twice 305? Of course it is, but is that the question? Why would we want to know whether Allentown, PA (area code 610), is twice Key West, FL (305)? Variables play different roles, and you can't tell a variable's role just by looking at it.

Some variables just tell us what group or category each individual belongs to. Are you male or female? Pierced or not? . . . What kinds of things can we learn about variables like these? A natural start is to *count* how many cases belong in each category. (Are you listening to music while reading this? We could count

[2] In database management, this kind of information is called "metadata."

It is wise to be careful. The *What* and *Why* of area codes are not as simple as they may first seem. When area codes were first introduced, 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. Now that phones have push-buttons, area codes have finally become just categories.

By international agreement, the International System of Units links together all systems of weights and measures. There are seven base units from which all other physical units are derived:

• Distance	Meter
• Mass	Kilogram
• Time	Second
• Electric current	Ampere
• Temperature	°Kelvin
• Amount of substance	Mole
• Intensity of light	Candela

A S ***Activity:* Recognize variables measured in a variety of ways.** This activity shows examples of the many ways to measure data.

A S ***Activities:* Variables.** Several activities show you how to begin working with data in your statistics package.

the number of students in the class who were and the number who weren't.) We'll look for ways to compare and contrast the sizes of such categories.

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

What kinds of things can we learn about measured variables? We can do a lot more than just counting categories. We can look for patterns and trends. (How much did you pay for your last movie ticket? What is the range of ticket prices available in your town? How has the price of a ticket changed over the past 20 years?)

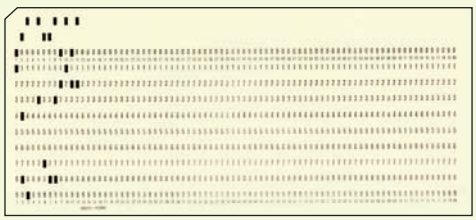
When a variable names categories and answers questions about how cases fall into those categories, we call it a **categorical variable.**[3] When a measured variable with units answers questions about the quantity of what is measured, we call it a **quantitative variable.** These types can help us decide what to do with a variable, but they are really more about what we hope to learn from a variable than about the variable itself. It's the questions we ask a variable (the *Why* of our analysis) that shape how we think about it and how we treat it.

Some variables can answer questions only about categories. If the values of a variable are words rather than numbers, it's a good bet that it is categorical. But some variables can answer both kinds of questions. Amazon could ask for your *Age* in years. That seems quantitative, and would be if the company wanted to know the average age of those customers who visit their site after 3 a.m. But suppose Amazon wants to decide which CD to offer you in a special deal—one by Raffi, Blink-182, Carly Simon, or Mantovani—and needs to be sure to have adequate supplies on hand to meet the demand. Then thinking of your age in one of the categories—child, teen, adult, or senior—might be more useful. If it isn't clear whether a variable is categorical or quantitative, think about *Why* you are looking at it and what you want it to tell you.

A typical course evaluation survey asks, "How valuable do you think this course will be to you?": 1 = Worthless; 2 = Slightly; 3 = Middling; 4 = Reasonably; 5 = Invaluable. Is *Educational Value* categorical or quantitative? Once again, we'll look to the *Why.* A teacher might just count the number of students who gave each response for her course, treating *Educational Value* as a categorical variable. When she wants to see whether the course is improving, she might treat the responses as the *amount* of perceived value—in effect, treating the variable as quantitative. But what are the units? There is certainly an *order* of perceived worth: Higher numbers indicate higher perceived worth. A course that averages 4.5 seems more valuable than one that averages 2, but we should be careful about treating *Educational Value* as

[3] You may also see it called a *qualitative variable.*

One tradition that hangs on in some quarters is to name variables with cryptic abbreviations written in uppercase letters. This can be traced back to the 1960s, when the very first statistics computer programs were controlled with instructions punched on cards. The earliest punch card equipment used only uppercase letters, and the earliest statistics programs limited variable names to six or eight characters, so variables were called things like PRSRF3. Modern programs do not have such restrictive limits, so there is no reason for variable names that you wouldn't use in an ordinary sentence.



purely quantitative. To treat it as quantitative, she'll have to imagine that it has "educational value units" or some similar arbitrary construction. Because there are no natural units, she should be cautious. Variables like this that report order without natural units are often called "ordinal" variables. But saying "that's an ordinal variable" doesn't get you off the hook. You must still look to the *Why* of your study to decide whether to treat it as categorical or quantitative.

FOR EXAMPLE **Identifying "What" and "Why" of HDTVs.**

Recap: A *Consumer Reports* article about 98 HDTVs lists each set's manufacturer, cost, screen size, type (LCD, plasma, or rear projection), and overall performance score (0–100).

Question: Are these variables categorical or quantitative? Include units where appropriate, and describe the "Why" of this investigation.

The "what" of this article includes the following variables:

- manufacturer (categorical);
- cost (in dollars, quantitative);
- screen size (in inches, quantitative);
- type (categorical);
- performance score (quantitative).

The magazine hopes to help consumers pick a good HDTV set.

Counts Count

In Statistics, we often count things. When Amazon considers a special offer of free shipping to customers, it might first analyze how purchases are shipped. They'd probably start by counting the number of purchases shipped by ground transportation, by second-day air, and by overnight air. Counting is a natural way to summarize the categorical variable *Shipping Method*. So every time we see counts, does that mean the variable is categorical? Actually, no.

We also use counts to measure the amounts of things. How many songs are on your digital music player? How many classes are you taking this semester? To measure these quantities, we'd naturally count. The variables (*Songs*, *Classes*) would be quantitative, and we'd consider the units to be "number of . . ." or, generically, just "counts" for short.

So we use counts in two different ways. When we count the cases in each category of a categorical variable, the category labels are the *What* and the individuals counted are the *Who* of our data. The counts themselves are not the

A S **Activity: Collect data in an experiment on yourself.** With the computer, you can experiment on yourself and then save the data. Go on to the subsequent related activities to check your understanding.

data, but are something we summarize about the data. Amazon counts the number of purchases in each category of the categorical variable *Shipping Method*. For this purpose (the *Why*), the *What* is shipping method and the *Who* is purchases.

Shipping Method	Number of Purchases
Ground	20,345
Second-day	7,890
Overnight	5,432

Other times our focus is on the amount of something, which we measure by counting. Amazon might record the number of teenage customers visiting their site each month to track customer growth and forecast CD sales (the *Why*). Now the *What* is *Teens*, the *Who* is *Months*, and the units are *Number of Teenage Customers*. *Teen* was a category when we looked at the categorical variable *Age*. But now it is a quantitative variable in its own right whose amount is measured by counting the number of customers.

Month	Number of Teenage Customers
January	123,456
February	234,567
March	345,678
April	456,789
May	. . .
. . .	. . .

Identifying Identifiers

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

Identifier variables themselves don't tell us anything useful about the categories because we know there is exactly one individual in each. However, they are crucial in this age of large data sets. They make it possible to combine data from different sources, to protect confidentiality, and to provide unique labels. The variables *UPS Tracking Number*, *Social Security Number*, and Amazon's *ASIN* are all examples of identifier variables.

You'll want to recognize when a variable is playing the role of an identifier so you won't be tempted to analyze it. There's probably a list of unique ID numbers for students in a class (so they'll each get their own grade confidentially), but you might worry about the professor who keeps track of the average of these numbers from class to class. Even though this year's average ID number happens to be higher than last's, it doesn't mean that the students are better.

Where, When, and How

A S **Self-Test: Review concepts about data.** Like the Just Checking sections of this textbook, but interactive. (Usually, we won't reference the *ActivStats* self-tests here, but look for one whenever you'd like to check your understanding or review material.)

We must know *Who*, *What*, and *Why* to analyze data. Without knowing these three, we don't have enough to start. Of course, we'd always like to know more. 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.

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. 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 *Think* step of any analysis. Make sure you know all three before you proceed to *Show* or *Tell* anything about the data.

JUST CHECKING

In the 2003 Tour de France, Lance Armstrong averaged 40.94 kilometers per hour (km/h) for the entire course, making it the fastest Tour de France in its 100-year history. In 2004, he made history again by winning the race for an unprecedented sixth time. In 2005, he became the only 7-time winner and once again set a new record for the fastest average speed. You can find data on all the Tour de France races on the DVD. Here are the first three and last ten lines of the data set. Keep in mind that the entire data set has nearly 100 entries.

1. List as many of the W's as you can for this data set.
2. Classify each variable as categorical or quantitative; if quantitative, identify the units.

Year	Winner	Country of origin	Total time (h/min/s)	Avg. speed (km/h)	Stages	Total distance ridden (km)	Starting riders	Finishing riders
1903	Maurice Garin	France	94.33.00	25.3	6	2428	60	21
1904	Henri Cornet	France	96.05.00	24.3	6	2388	88	23
1905	Louis Trousselier	France	112.18.09	27.3	11	2975	60	24
⋮								
1999	Lance Armstrong	USA	91.32.16	40.30	20	3687	180	141
2000	Lance Armstrong	USA	92.33.08	39.56	21	3662	180	128
2001	Lance Armstrong	USA	86.17.28	40.02	20	3453	189	144
2002	Lance Armstrong	USA	82.05.12	39.93	20	3278	189	153
2003	Lance Armstrong	USA	83.41.12	40.94	20	3427	189	147
2004	Lance Armstrong	USA	83.36.02	40.53	20	3391	188	147
2005	Lance Armstrong	USA	86.15.02	41.65	21	3608	189	155
2006	Óscar Periero	Spain	89.40.27	40.78	20	3657	176	139
2007	Alberto Contador	Spain	91.00.26	38.97	20	3547	189	141
2008	Carlos Sastre	Spain	87.52.52	40.50	21	3559	199	145

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 Web site. We'll sometimes suggest search terms and offer other guidance.

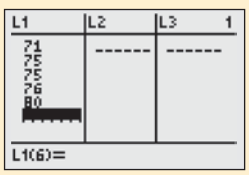
Some words of caution, though: Data found on Internet sites may not be formatted in the best way for use in statistics software. Although you may see a data table in standard form, an attempt to copy the data may leave you with a single column of values. you may have to work in your favorite statistics 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.

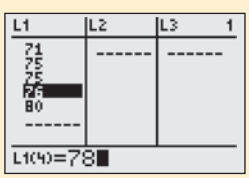
WHAT CAN GO WRONG?

- **Don't label a variable as categorical or quantitative without thinking about the question you want it to answer.** The same variable can sometimes take on different roles.
- **Just because your variable's values are numbers, don't assume that it's quantitative.** 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 Web site. The question that respondents answered may have been posed in a way that influenced their responses.

TI Tips

Working with data





You'll need to be able to enter and edit data in your calculator. Here's how.

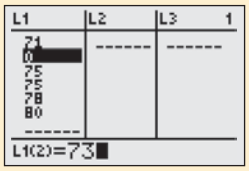
To enter data:

Hit the STAT button, and choose EDIT from the menu. You'll see a set of columns labeled L1, L2, and so on. Here is where you can enter, change, or delete a set of data.

Let's enter the heights (in inches) of the five starting players on a basketball team: 71, 75, 75, 76, and 80. Move the cursor to the space under L1, type in 71, and hit ENTER (or the down arrow). There's the first player. Now enter the data for the rest of the team.

To change a datum:

Suppose the 76" player grew since last season; his height should be listed as 78". Use the arrow keys to move the cursor onto the 76, then change the value and ENTER the correction.



To add more data:

We want to include the sixth man, 73" tall. It would be easy to simply add this new datum to the end of the list. However, sometimes the order of the data matters, so let's place this datum in numerical order. Move the cursor to the desired position (atop the first 75). Hit 2ND INS, then ENTER the 73 in the new space.

To delete a datum:

The 78" player just quit the team. Move the cursor there. Hit DEL. Bye.

To clear the datalist:

Finished playing basketball? Move the cursor atop the L1. Hit CLEAR, then ENTER (or down arrow). You should now have a blank datalist, ready for you to enter your next set of values.

Lost a datalist?

Oops! Is L1 now missing entirely? Did you delete L1 by mistake, instead of just *clearing* it? Easy problem to fix: buy a new calculator. No? OK, then simply go to the STAT EDIT menu, and run SetUpEditor to recreate all the lists.

WHAT HAVE WE LEARNED?

We've learned that data are information in a context.

- The W's help nail down the context: *Who, What, Why, Where, When,* and *hoW.*
- We must know at least the *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.

We treat variables in two basic ways: as *categorical* or *quantitative*.

- Categorical variables identify a category for each case. Usually, we think about the counts of cases that fall into 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 treat a variable as categorical or quantitative depending on what we want to learn from it, which means that some variables can't be pigeonholed as one type or the other. That's an early hint that in Statistics we can't always pin things down precisely.

Terms

Context 8. 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.

Data 8. Systematically recorded information, whether numbers or labels, together with its context.

Data table 8. An arrangement of data in which each row represents a case and each column represents a variable.

Case 9. A case is an individual about whom or which we have data.

Population 9. All the cases we wish we knew about.

Sample 9. The cases we actually examine in seeking to understand the much larger population.

Variable 9. A variable holds information about the same characteristic for many cases.

Units 10. A quantity or amount adopted as a standard of measurement, such as dollars, hours, or grams.

Categorical variable 10. A variable that names categories (whether with words or numerals) is called categorical.

Quantitative variable 10. A variable in which the numbers act as numerical values is called quantitative. Quantitative variables always have units.

Skills

THINK

- Be able to identify the *Who, What, When, Where, Why,* and *How* of data, or recognize when some of this information has not been provided.
- Be able to identify the cases and variables in any data set.
- Be able to identify the population from which a sample was chosen.
- Be able to classify a variable as categorical or quantitative, depending on its use.
- For any quantitative variable, be able to identify the units in which the variable has been measured (or note that they have not been provided).

TELL

- Be able to describe a variable in terms of its *Who, What, When, Where, Why,* and *How* (and be prepared to remark when that information is not provided).

DATA ON THE COMPUTER

A S

***Activity:* Examine the Data.** Take a look at your own data from your experiment (p. 12) and get comfortable with your statistics package as you find out about the experiment test results.

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.

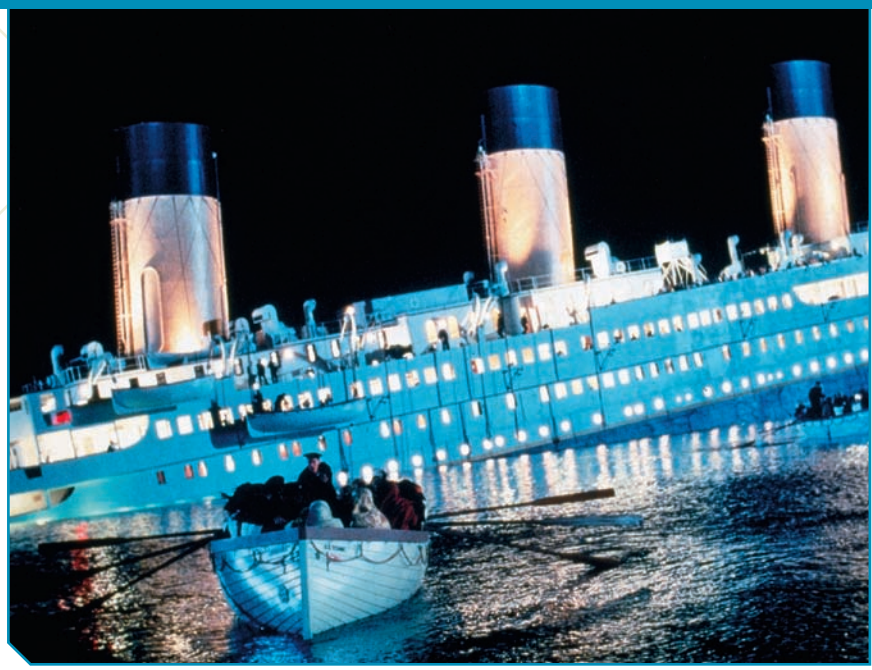
EXERCISES

1. **Voters.** A February 2007 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?
2. **Mood.** A January 2007 Gallup Poll question asked, "In general, do you think things have gotten better or gotten worse in this country in the last five years?" Possible answers were "Better", "Worse", "No Change", "Don't Know", and "No Response". What kind of variable is the response?
3. **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?
4. **Stress.** A medical researcher measures the increase in heart rate of patients under a stress test. What kind of variable is the researcher studying?

(Exercises 5–12) *For each description of data, identify Who and What were investigated and the population of interest.*

CHAPTER 3

Displaying and Describing Categorical Data



WHO	People on the *Titanic*
WHAT	Survival status, age, sex, ticket class
WHEN	April 14, 1912
WHERE	North Atlantic
HOW	A variety of sources and Internet sites
WHY	Historical interest

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 a.m., the *Titanic*, thought by many to be unsinkable, had sunk, leaving more than 1500 passengers and crew members on board to meet their icy fate.

Here are 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 *Survival* status (Dead or Alive), *Age* (Adult or Child), *Sex* (Male or Female), and ticket *Class* (First, Second, Third, or Crew).

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.

AS ***Video:* The Incident** tells the story of the *Titanic*, and includes rare film footage.

Survival	Age	Sex	Class
Dead	Adult	Male	Third
Dead	Adult	Male	Crew
Dead	Adult	Male	Third
Dead	Adult	Male	Crew
Dead	Adult	Male	Crew
Dead	Adult	Male	Crew
Alive	Adult	Female	First
Dead	Adult	Male	Third
Dead	Adult	Male	Crew

Table 3.1
Part of a data table showing four variables for nine people aboard the *Titanic*.

The Three Rules of Data Analysis

So, what should we do with data like these? There are three things you should always do first with data:

1. **Make a picture.** A display of your data will reveal things you are not likely to see in a table of numbers and will help you to *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. A picture will also show you the things you did not expect to see: the 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.

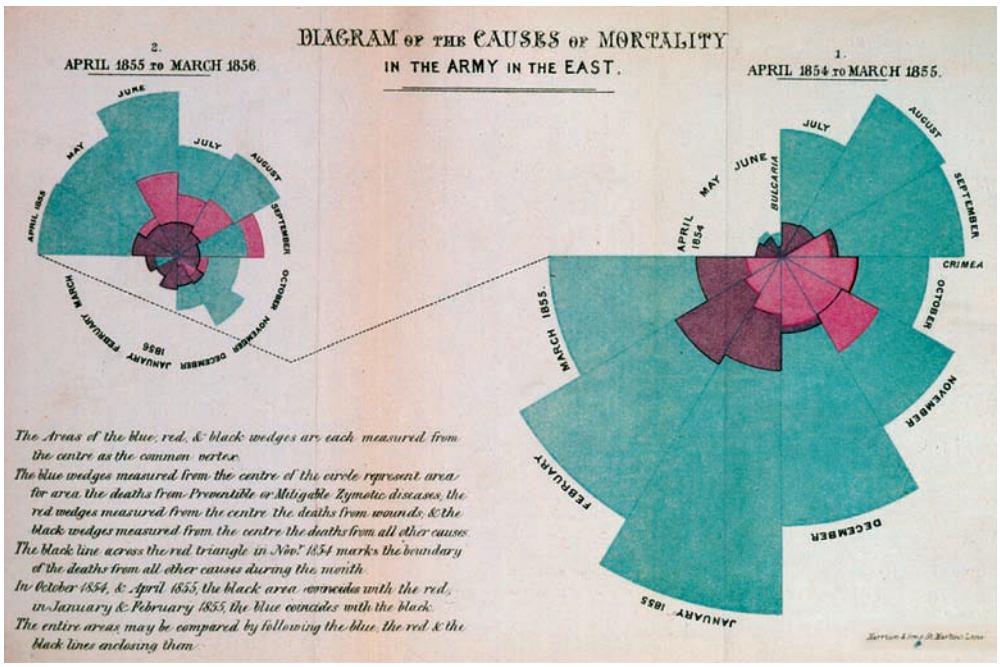
These are the three rules of data analysis. There are pictures of data throughout the book, and new kinds keep showing up. These days, technology makes drawing pictures of data easy, so there is no reason not to follow the three rules.

FIGURE 3.1 ***A Picture to Tell a Story***

Florence Nightingale (1820–1910), a founder of modern nursing, was also a pioneer in health management, statistics, and epidemiology. She was the first female member of the British Statistical Society and was granted honorary membership in the newly formed American Statistical Association.

To argue forcefully for better hospital conditions for soldiers, she and her colleague, Dr. William Farr, invented this display, which showed that in the Crimean War, far more soldiers died of illness and infection than of battle wounds. Her campaign succeeded in improving hospital conditions and nursing for soldiers.

Florence Nightingale went on to apply statistical methods to a variety of important health issues and published more than 200 books, reports, and pamphlets during her long and illustrious career.



Frequency Tables: Making Piles

A S ***Activity:*** **Make and examine a table of counts.** Even data on something as simple as hair color can reveal surprises when you organize it in a data table.

To make a picture of data, the first thing we have to do is to make piles. Making piles is the beginning of understanding about data. We pile together things that seem to go together, so we can see how the cases distribute across different categories. For categorical data, piling is easy. We just count the number of cases corresponding to each category and pile them up.

One way to put all 2201 people on the *Titanic* into piles is by ticket *Class*, counting up how many had each kind of ticket. We can organize these counts into a **frequency table,** which records the totals and the category names.

Even when we have thousands of cases, a variable like ticket *Class*, with only a few categories, has a frequency table that's easy to read. A frequency table with dozens or hundreds of categories would be much harder to read. We use the names of the categories to label each row in the frequency table. For ticket *Class*, these are "First," "Second," "Third," and "Crew."

Class	Count
First	325
Second	285
Third	706
Crew	885

Table 3.2

A frequency table of the *Titanic* passengers.

Class	%
First	14.77
Second	12.95
Third	32.08
Crew	40.21

Table 3.3
A relative frequency table for the same data.

Counts are useful, but sometimes we want to know the fraction or **proportion** of the data in each category, so we divide the counts by the total number of cases. Usually we multiply by 100 to express these proportions as **percentages**. A **relative frequency table** displays the *percentages*, rather than the counts, of the values in each category. Both types of tables show how the cases are distributed across the categories. In this way, they describe the **distribution** of a categorical variable because they name the possible categories and tell how frequently each occurs.

The Area Principle

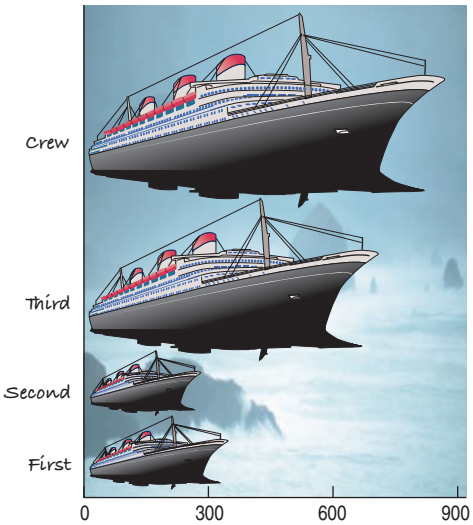


FIGURE 3.2
How many people were in each class on the Titanic? From this display, it looks as though the service must have been great, since most aboard were crew members. Although the length of each ship here corresponds to the correct number, the impression is all wrong. In fact, only about 40% were crew.

Now that we have the frequency table, we're ready to follow the three rules of data analysis and make a picture of the data. But a bad picture can distort our understanding rather than help it. Here's a graph of the *Titanic* data. What impression do you get about who was aboard the ship?

It sure looks like most of the people on the *Titanic* were crew members, with a few passengers along for the ride. That doesn't seem right. What's wrong? The lengths of the ships *do* match the totals in the table. (You can check the scale at the bottom.) However, experience and psychological tests show that 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. Since there were about 3 times as many crew as second-class passengers, the ship depicting the number of crew is about 3 times longer than the ship depicting second-class passengers, but it occupies about 9 times the area. As you can see from the frequency table (Table 3.2), 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.

Bar Charts

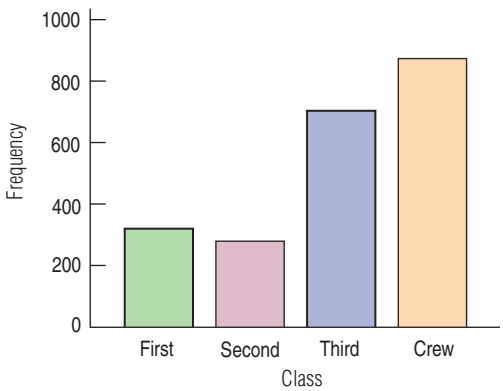
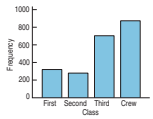
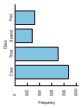


FIGURE 3.3 ***People on the* Titanic *by Ticket Class***
With the area principle satisfied, we can see the true distribution more clearly.

Here's a chart that obeys the area principle. It's not as visually entertaining as the ships, but it does give an *accurate* visual impression of the distribution. The height of each bar shows the count for its category. The bars are the same width, so their heights determine their areas, and the areas are proportional to the counts in each class. Now it's easy to see that the majority of people on board were *not* crew, as the ships picture led us to believe. We can also see that there were about 3 times as many crew 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.

A **bar chart** displays the distribution of a categorical variable, showing the counts for each category next to each other for easy comparison. Bar charts should have small spaces between the bars to indicate that these are freestanding bars that could be rearranged into any order. The bars are lined up along a common base.

Usually they stick up like this  but sometimes they run sideways like this 

If we really want to draw attention to the relative *proportion* of passengers falling into each of these classes, we could replace the counts with percentages and use a **relative frequency bar chart.**

A S ***Activity:*** **Bar Charts.** Watch bar charts grow from data; then use your statistics package to create some bar charts for yourself.

For some reason, some computer programs give the name "bar chart" to any graph that uses bars. And others use different names according to whether the bars are horizontal or vertical. Don't be misled. "Bar chart" is the term for a *display of counts of a categorical variable* with bars.

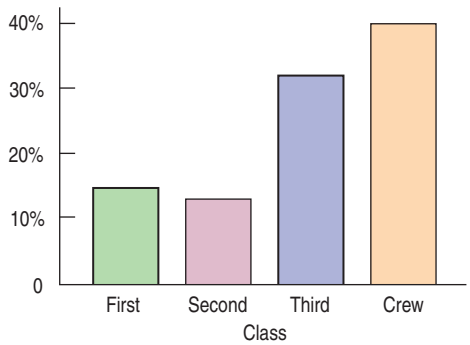


FIGURE 3.4
The relative frequency bar chart looks the same as the bar chart (Figure 3.3) but shows the proportion of people in each category rather than the counts.

Pie Charts

Another common display that shows how a whole group breaks into several categories is a pie chart. **Pie charts** show the whole group of cases as a circle. They slice the circle into pieces whose sizes are proportional to the fraction of the whole in each category.

Pie charts give a quick impression of how a whole group is partitioned into smaller groups. Because we're used to cutting up pies into 2, 4, or 8 pieces, pie charts are good for seeing relative frequencies near 1/2, 1/4, or 1/8. For example, you may be able to tell that the pink slice, representing the second-class passengers, is very close to 1/8 of the total. It's harder to see that there were about twice as many third-class as first-class passengers. Which category had the most passengers? Were there more crew or more third-class passengers? Comparisons such as these are easier in a bar chart.

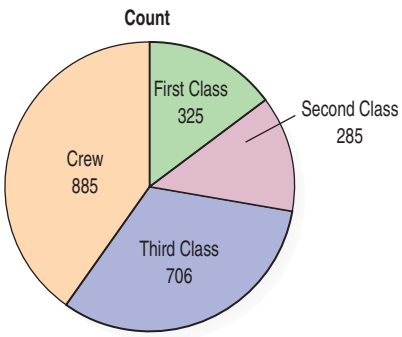


FIGURE 3.5 ***Number of*** **Titanic** ***passengers in each class***

Think before you draw. Our first rule of data analysis is *Make a picture*. But what kind of picture? We don't have a lot of options—yet. There's more to Statistics than pie charts and bar charts, and knowing when to use each type of graph is a critical first step in data analysis. That decision depends in part on what type of data we have.

It's important to check that the data are appropriate for whatever method of analysis you choose. Before you make a bar chart or a pie chart, always check the

Categorical Data Condition: The data are counts or percentages of individuals in categories.

If you want to make a relative frequency bar chart or a pie chart, you'll need to also make sure that the categories don't overlap so that no individual is counted twice. If the categories do overlap, you can still make a bar chart, but the percentages won't add up to 100%. For the *Titanic* data, either kind of display is appropriate because the categories don't overlap.

Throughout this course, you'll see that doing Statistics right means selecting the proper methods. That means you have to *Think* about the situation at hand. An important first step, then, is to check that the type of analysis you plan is appropriate. The Categorical Data Condition is just the first of many such checks.

Contingency Tables: Children and First-Class Ticket Holders First?

A S ***Activity:* Children at Risk.** This activity looks at the fates of children aboard the *Titanic*; the subsequent activity shows how to make such tables on a computer.

We know how many tickets of each class were sold on the *Titanic,* and we know that only about 32% of all those aboard the *Titanic* survived. After looking at the distribution of each variable by itself, it's natural and more interesting to ask how they relate. Was there a relationship between the kind of ticket a passenger held and the passenger's chances of making it into the lifeboat? To answer this question, we need to look at the two categorical variables *Class* and *Survival* together.

To look at two categorical variables together, we often arrange the counts in a two-way table. Here is a two-way table of those aboard the *Titanic,* classified according to the class of ticket and whether the ticket holder survived or didn't. Because the table shows how the individuals are distributed along each variable, contingent on the value of the other variable, such a table is called a **contingency table.**

Contingency table of ticket *Class* and *Survival.* The bottom line of "Totals" is the same as the previous frequency table.
Table 3.4

		Class				
		First	Second	Third	Crew	Total
Survival	Alive	203	118	178	212	**711**
	Dead	122	167	528	673	**1490**
	Total	**325**	**285**	**706**	**885**	**2201**

A bell-shaped artifact from the Titanic.

The margins of the table, both on the right and at the bottom, give totals. The bottom line of the table is just the frequency distribution of ticket *Class.* The right column of the table is the frequency distribution of the variable *Survival.* When presented like this, in the margins of a contingency table, the frequency distribution of one of the variables is called its **marginal distribution.**

Each **cell** of the table gives the count for a combination of values of the two variables. If you look down the column for second-class passengers to the first cell, you can see that 118 second-class passengers survived. Looking at the third-class passengers, you can see that more third-class passengers (178) survived. Were second-class passengers more likely to survive? Questions like this are easier to address by using percentages. The 118 survivors in second class were 41.4% of the total 285 second-class passengers, while the 178 surviving third-class passengers were only 25.2% of that class's total.

We know that 118 second-class passengers survived. We could display this number as a percentage—but as a percentage of what? The total number of passengers? (118 is 5.4% of the total: 2201.) The number of second-class passengers?

(118 is 41.4% of the 285 second-class passengers.) The number of survivors? (118 is 16.6% of the 711 survivors.) All of these are possibilities, and all are potentially useful or interesting. You'll probably wind up calculating (or letting your technology calculate) lots of percentages. Most statistics programs offer a choice of total percent, row percent, or column percent for contingency tables. Unfortunately, they often put them all together with several numbers in each cell of the table. The resulting table holds lots of information, but it can be hard to understand:

Another contingency table of ticket Class. This time we see not only the counts for each combination of *Class* and *Survival* (in bold) but the percentages these counts represent. For each count, there are three choices for the percentage: by row, by column, and by table total. There's probably too much information here for this table to be useful.

Table 3.5

			Class				
			First	Second	Third	Crew	Total
Survival	Alive	Count	203	118	178	212	711
		% of Row	28.6%	16.6%	25.0%	29.8%	100%
		% of Column	62.5%	41.4%	25.2%	24.0%	32.3%
		% of Table	9.2%	5.4%	8.1%	9.6%	32.3%
	Dead	Count	122	167	528	673	1490
		% of Row	8.2%	11.2%	35.4%	45.2%	100%
		% of Column	37.5%	58.6%	74.8%	76.0%	67.7%
		% of Table	5.6%	7.6%	24.0%	30.6%	67.7%
	Total	Count	325	285	706	885	2201
		%of Row	14.8%	12.9%	32.1%	40.2%	100%
		% of Column	100%	100%	100%	100%	100%
		% of Table	14.8%	12.9%	32.1%	40.2%	100%

To simplify the table, let's first pull out the percent of table values:

A contingency table of *Class* by *Survival* with only the table percentages

Table 3.6

		Class				
		First	Second	Third	Crew	Total
Survival	Alive	9.2%	5.4%	8.1%	9.6%	32.3%
	Dead	5.6%	7.6%	24.0%	30.6%	67.7%
	Total	14.8%	12.9%	32.1%	40.2%	100%

These percentages tell us what percent of *all* passengers belong to each combination of column and row category. For example, we see that although 8.1% of the people aboard the *Titanic* were surviving third-class ticket holders, only 5.4% were surviving second-class ticket holders. Is this fact useful? Comparing these percentages, you might think that the chances of surviving were better in third class than in second. But be careful. There were many more third-class than second-class passengers on the *Titanic,* so there were more third-class survivors. That group is a larger percentage of the passengers, but is that really what we want to know?

Percent of what? The English language can be tricky when we talk about percentages. If you're asked "What percent *of the survivors* were in second class?" it's pretty clear that we're interested only in survivors. It's as if we're restricting the *Who* in the question to the survivors, so we should look at the number of second-class passengers among all the survivors—in other words, the row percent.

But if you're asked "What percent were second-class passengers who survived?" you have a different question. Be careful; here, the *Who* is everyone on board, so 2201 should be the denominator, and the answer is the table percent.

And if you're asked "What percent of the second-class passengers survived?" you have a third question. Now the *Who* is the second-class passengers, so the denominator is the 285 second-class passengers, and the answer is the column percent.

Always be sure to ask "percent of what?" That will help you to know the *Who* and whether we want *row*, *column*, or *table* percentages.

FOR EXAMPLE Finding marginal distributions

In January 2007, a Gallup poll asked 1008 Americans age 18 and over whether they planned to watch the upcoming Super Bowl. The pollster also asked those who planned to watch whether they were looking forward more to seeing the football game or the commercials. The results are summarized in the table:

Response	Sex: Male	Female	Total
Game	279	200	479
Commercials	81	156	237
Won't watch	132	160	292
Total	492	516	1008

Question: What's the marginal distribution of the responses?

To determine the percentages for the three responses, divide the count for each response by the total number of people polled:

$$\frac{479}{1008} = 47.5\% \quad \frac{237}{1008} = 23.5\% \quad \frac{292}{1008} = 29.0\%$$

According to the poll, 47.5% of American adults were looking forward to watching the Super Bowl game, 23.5% were looking forward to watching the commercials, and 29% didn't plan to watch at all.

Conditional Distributions

The more interesting questions are *contingent*. We'd like to know, for example, what percentage of *second-class passengers* survived and how that compares with the survival rate for third-class passengers.

It's more interesting to ask whether the chance of surviving the *Titanic* sinking *depended* on ticket class. We can look at this question in two ways. First, we could ask how the distribution of ticket *Class* changes between survivors and nonsurvivors. To do that, we look at the *row percentages:*

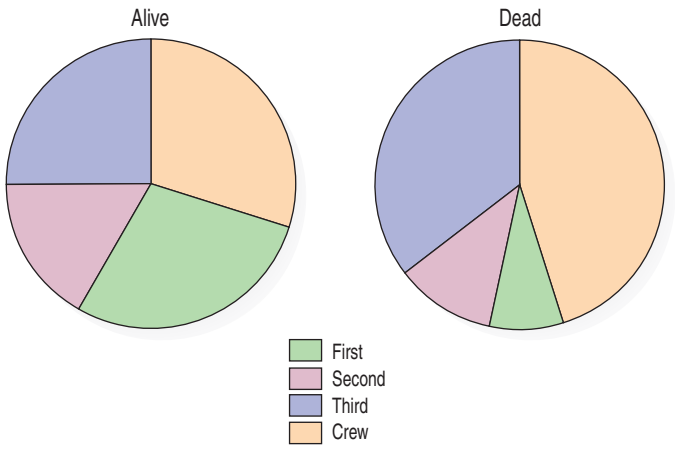
Table 3.7 **The conditional distribution of ticket *Class*** conditioned on each value of *Survival: Alive* and *Dead*.

Survival	Class: First	Second	Third	Crew	Total
Alive	203 28.6%	118 16.6%	178 25.0%	212 29.8%	711 100%
Dead	122 8.2%	167 11.2%	528 35.4%	673 45.2%	1490 100%

By focusing on each row separately, we see the distribution of class under the *condition* of surviving or not. The sum of the percentages in each row is 100%, and we divide that up by ticket class. In effect, we temporarily restrict the *Who* first to survivors and make a pie chart for them. Then we refocus the *Who* on the nonsurvivors and make their pie chart. These pie charts show the distribution of ticket classes *for each row* of the table: survivors and nonsurvivors. The distributions we create this way are called **conditional distributions,** because they show the distribution of one variable for just those cases that satisfy a condition on another variable.

FIGURE 3.6

Pie charts of the conditional distributions of ticket Class *for the survivors and nonsurvivors, separately. Do the distributions appear to be the same? We're primarily concerned with percentages here, so pie charts are a reasonable choice.*



FOR EXAMPLE Finding conditional distributions

Recap: The table shows results of a poll asking adults whether they were looking forward to the Super Bowl game, looking forward to the commercials, or didn't plan to watch.

Question: How do the conditional distributions of interest in the commercials differ for men and women?

Response	Sex: Male	Female	Total
Game	279	200	479
Commercials	81	156	237
Won't watch	132	160	292
Total	492	516	1008

Look at the group of people who responded "Commercials" and determine what percent of them were male and female:

$$\frac{81}{237} = 34.2\% \quad \frac{156}{237} = 65.8\%$$

Women make up a sizable majority of the adult Americans who look forward to seeing Super Bowl commercials more than the game itself. Nearly 66% of people who voiced a preference for the commercials were women, and only 34% were men.

But we can also turn the question around. We can look at the distribution of *Survival* for each category of ticket *Class*. To do this, we look at the *column percentages*. Those show us whether the chance of surviving was roughly the same *for each of the four classes*. Now the percentages in each column add to 100%, because we've restricted the *Who*, in turn, to each of the four ticket classes:

A contingency table of *Class* by *Survival* with only counts and column percentages. Each column represents the conditional distribution of *Survival* for a given category of ticket *Class*.

Table 3.8

Survival		Class: First	Second	Third	Crew	Total
Alive	Count	203	118	178	212	711
	% of Column	62.5%	41.4%	25.2%	24.0%	32.3%
Dead	Count	122	167	528	673	1490
	% of Column	37.5%	58.6%	74.8%	76.0%	67.7%
Total	Count	325	285	706	885	2201
		100%	100%	100%	100%	100%

Looking at how the percentages change across each row, it sure looks like ticket class mattered in whether a passenger survived. To make it more vivid, we could show the distribution of *Survival* for each ticket class in a display. Here's a side-by-side bar chart showing percentages of surviving and not for each category:

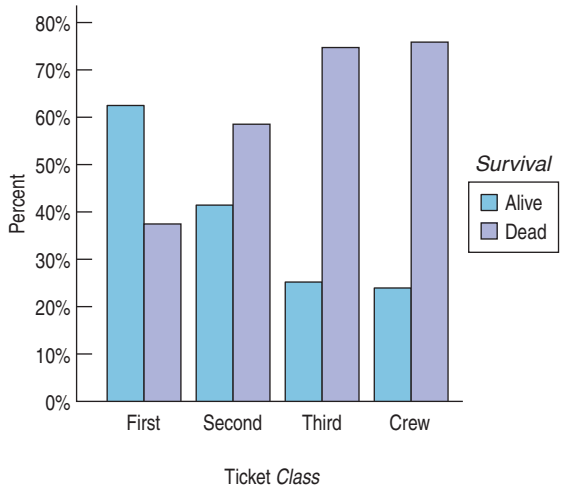


FIGURE 3.7
Side-by-side bar chart *showing the conditional distribution of Survival for each category of ticket Class. The corresponding pie charts would have only two categories in each of four pies, so bar charts seem the better alternative.*

These bar charts are simple because, for the variable *Survival*, we have only two alternatives: Alive and Dead. When we have only two categories, we really need to know only the percentage of one of them. Knowing the percentage that survived tells us the percentage that died. We can use this fact to simplify the display even more by dropping one category. Here are the percentages of dying *across the classes* displayed in one chart:

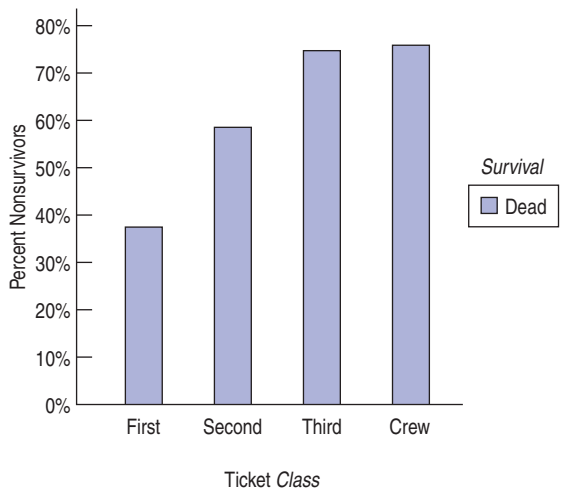


FIGURE 3.8
Bar chart showing just nonsurvivor percentages for each value of ticket Class. Because we have only two values, the second bar doesn't add any information. Compare this chart to the side-by-side bar chart shown earlier.

Now it's easy to compare the risks. Among first-class passengers, 37.5% perished, compared to 58.6% for second-class ticket holders, 74.8% for those in third class, and 76.0% for crew members.

If the risk had been about the same across the ticket classes, we would have said that survival was *independent* of class. But it's not. The differences we see among these conditional distributions suggest that survival may have depended on ticket class. You may find it useful to consider conditioning on each variable in a contingency table in order to explore the dependence between them.

TI-*nspire*

Conditional distributions and association. Explore the *Titanic* data to see which passengers were most likely to survive.

It is interesting to know that *Class* and *Survival* are associated. That's an important part of the *Titanic* story. And we know how important this is because the margins show us the actual numbers of people involved.

Variables can be associated in many ways and to different degrees. The best way to tell whether two variables are associated is to ask whether they are *not*.[1] In a contingency table, when the distribution of *one* variable is the same for all categories of another, we say that the variables are **independent**. That tells us there's no association between these variables. We'll see a way to check for independence formally later in the book. For now, we'll just compare the distributions.

FOR EXAMPLE Looking for associations between variables

Recap: The table shows results of a poll asking adults whether they were looking forward to the Super Bowl game, looking forward to the commercials, or didn't plan to watch.

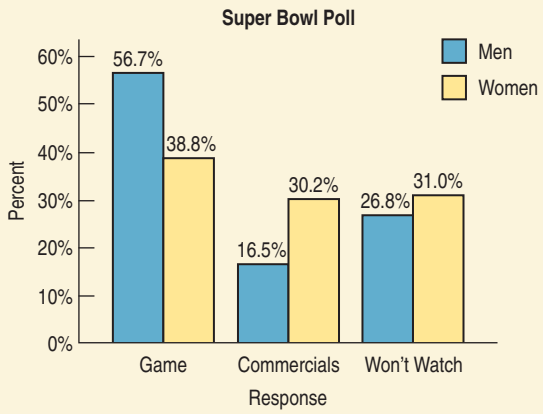
Question: Does it seem that there's an association between interest in Super Bowl TV coverage and a person's sex?

Response	Sex: Male	Sex: Female	Total
Game	279	200	479
Commercials	81	156	237
Won't watch	132	160	292
Total	492	516	1008

First find the distribution of the three responses for the men (the column percentages):

$$\frac{279}{492} = 56.7\% \qquad \frac{81}{492} = 16.5\% \qquad \frac{132}{492} = 26.8\%$$

Then do the same for the women who were polled, and display the two distributions with a side-by-side bar chart:



Based on this poll it appears that women were only slightly less interested than men in watching the Super Bowl telecast: 31% of the women said they didn't plan to watch, compared to just under 27% of men. Among those who planned to watch, however, there appears to be an association between the viewer's sex and what the viewer is most looking forward to. While more women are interested in the game (39%) than the commercials (30%), the margin among men is much wider: 57% of men said they were looking forward to seeing the game, compared to only 16.5% who cited the commercials.

[1] This kind of "backwards" reasoning shows up surprisingly often in science—and in Statistics. We'll see it again.

JUST CHECKING

A Statistics class reports the following data on Sex and Eye Color for students in the class:

		Eye Color			
		Blue	Brown	Green/Hazel/Other	Total
Sex	Males	6	20	6	32
	Females	4	16	12	32
	Total	10	36	18	64

1. What percent of females are brown-eyed?
2. What percent of brown-eyed students are female?
3. What percent of students are brown-eyed females?
4. What's the distribution of Eye Color?
5. What's the conditional distribution of Eye Color for the males?
6. Compare the percent who are female among the blue-eyed students to the percent of all students who are female.
7. Does it seem that Eye Color and Sex are independent? Explain.

Segmented Bar Charts

We could display the *Titanic* information by dividing up bars rather than circles. The resulting **segmented bar chart** treats each bar as the "whole" and divides it proportionally into segments corresponding to the percentage in each group. We can clearly see that the distributions of ticket *Class* are different, indicating again that survival was not independent of ticket *Class*.

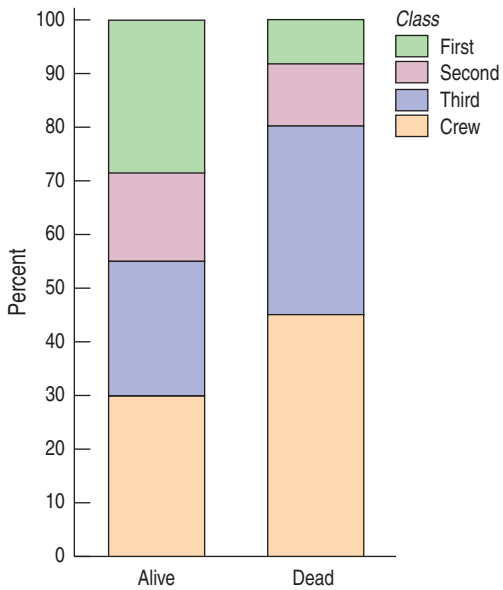


FIGURE 3.9 ***A segmented bar chart for Class by Survival***

Notice that although the totals for survivors and nonsurvivors are quite different, the bars are the same height because we have converted the numbers to percentages. *Compare this display with the side-by-side pie charts of the same data in Figure 3.6.*

STEP-BY-STEP EXAMPLE — Examining Contingency Tables

Medical researchers followed 6272 Swedish men for 30 years to see if there was any association between the amount of fish in their diet and prostate cancer ("Fatty Fish Consumption and Risk of Prostate Cancer," *Lancet*, June 2001). Their results are summarized in this table:

We asked for a picture of a man eating fish. This is what we got.

Fish Consumption	Prostate Cancer: No	Prostate Cancer: Yes
Never/seldom	110	14
Small part of diet	2420	201
Moderate part	2769	209
Large part	507	42

Table 3.9

Question: Is there an association between fish consumption and prostate cancer?

Plan Be sure to state what the problem is about.

I want to know if there is an association between fish consumption and prostate cancer.

Variables Identify the variables and report the W's.

The individuals are 6272 Swedish men followed by medical researchers for 30 years. The variables record their fish consumption and whether or not they were diagnosed with prostate cancer.

Be sure to check the appropriate condition.

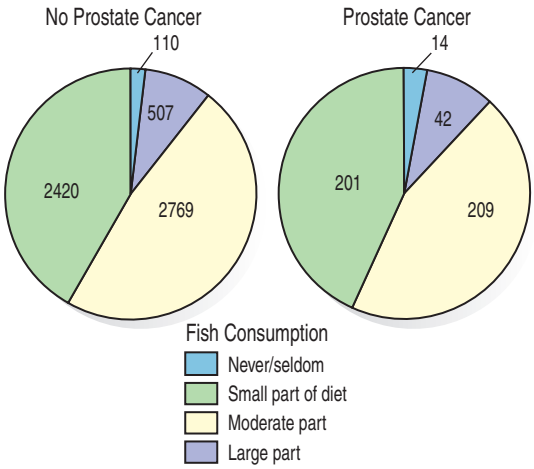
✓ ***Categorical Data Condition:*** I have counts for both fish consumption and cancer diagnosis. The categories of diet do not overlap, and the diagnoses do not overlap. It's okay to draw pie charts or bar charts.

Mechanics It's a good idea to check the marginal distributions first before looking at the two variables together.

Fish Consumption	Prostate Cancer: No	Prostate Cancer: Yes	Total
Never/seldom	110	14	**124 (2.0%)**
Small part of diet	2420	201	**2621 (41.8%)**
Moderate part	2769	209	**2978 (47.5%)**
Large part	507	42	**549 (8.8%)**
Total	**5806 (92.6%)**	**466 (7.4%)**	**6272 (100%)**

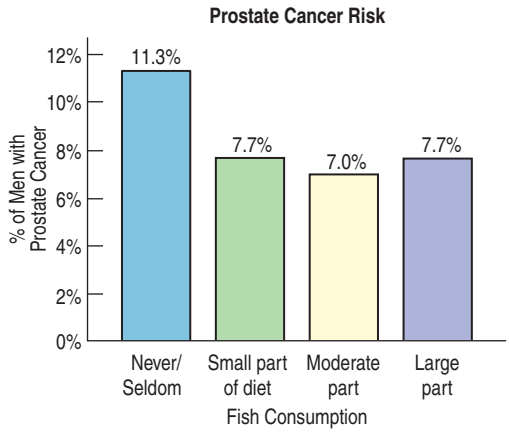
Two categories of the diet are quite small, with only 2.0% Never/Seldom eating fish and 8.8% in the "Large part" category. Overall, 7.4% of the men in this study had prostate cancer.

Then, make appropriate displays to see whether there is a difference in the relative proportions. These pie charts compare fish consumption for men who have prostate cancer to fish consumption for men who don't.



It's hard to see much difference in the pie charts. So, I made a display of the row percentages. Because there are only two alternatives, I chose to display the risk of prostate cancer for each group:

Both pie charts and bar charts can be used to compare conditional distributions. Here we compare prostate cancer rates based on differences in fish consumption.



Conclusion Interpret the patterns in the table and displays in context. If you can, discuss possible real-world consequences. Be careful not to overstate what you see. The results may not generalize to other situations.

Overall, there is a 7.4% rate of prostate cancer among men in this study. Most of the men (89.3%) ate fish either as a moderate or small part of their diet. From the pie charts, it's hard to see a difference in cancer rates among the groups. But in the bar chart, it looks like the cancer rate for those who never/seldom ate fish may be somewhat higher.

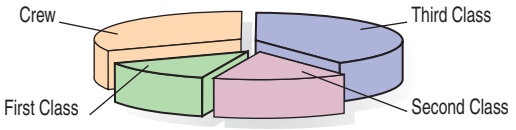
However, only 124 of the 6272 men in the study fell into this category, and only 14 of them developed prostate cancer. More study would probably be needed before we would recommend that men change their diets.[2]

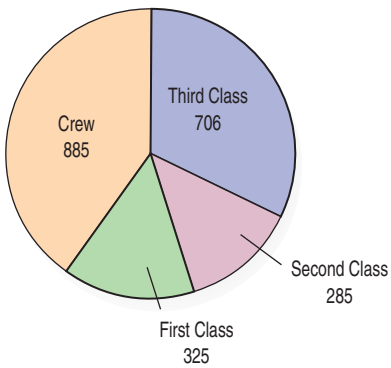
[2] The original study actually used pairs of twins, which enabled the researchers to discern that the risk of cancer for those who never ate fish actually *was* substantially greater. Using pairs is a special way of gathering data. We'll discuss such study design issues and how to analyze the data in the later chapters.

This study is an example of looking at a sample of data to learn something about a larger population. We care about more than these particular 6272 Swedish men. We hope that learning about their experiences will tell us something about the value of eating fish in general. That raises the interesting question of what population we think this sample might represent. Do we hope to learn about all Swedish men? About all men? About the value of eating fish for all adult humans? [3] Often, it can be hard to decide just which population our findings may tell us about, but that also is how researchers decide what to look into in future studies.

WHAT CAN GO WRONG?

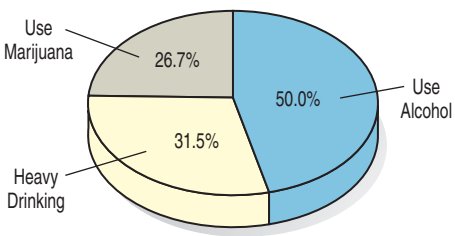
- **Don't violate the area principle.** This is probably the most common mistake in a graphical display. It is often made in the cause of artistic presentation. Here, for example, are two displays of the pie chart of the *Titanic* passengers by class:





The one on the left looks pretty, doesn't it? But showing the pie on a slant violates the area principle and makes it much more difficult to compare fractions of the whole made up of each class—the principal feature that a pie chart ought to show.

- **Keep it honest.** Here's a pie chart that displays data on the percentage of high school students who engage in specified dangerous behaviors as reported by the Centers for Disease Control. What's wrong with this plot?

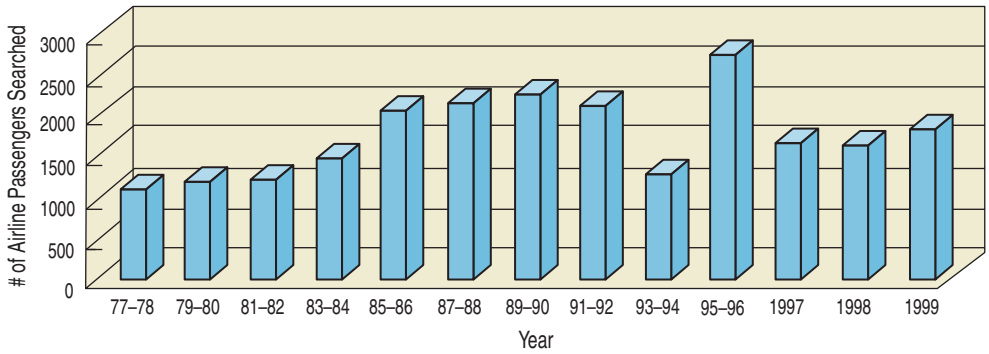


Try adding up the percentages. Or look at the 50% slice. Does it look right? Then think: What are these percentages of? Is there a "whole" that has been sliced up? In a pie chart, the proportions shown by each slice of the pie must add up to 100% and each individual must fall into only one category. Of course, showing the pie on a slant makes it even harder to detect the error.

(*continued*)

[3] Probably not, since we're looking only at prostate cancer risk.

Here's another. This bar chart shows the number of airline passengers searched in security screening, by year:



Looks like things didn't change much in the final years of the 20th century—until you read the bar labels and see that the last three bars represent single years while all the others are for *pairs* of years. Of course, the false depth makes it harder to see the problem.

- **Don't confuse similar-sounding percentages.** These percentages sound similar but are different:
  - The percentage of the passengers who were both in first class and survived: This would be 203/2201, or 9.4%.
  - The percentage of the first-class passengers who survived: This is 203/325, or 62.5%.
  - The percentage of the survivors who were in first class: This is 203/711, or 28.6%.

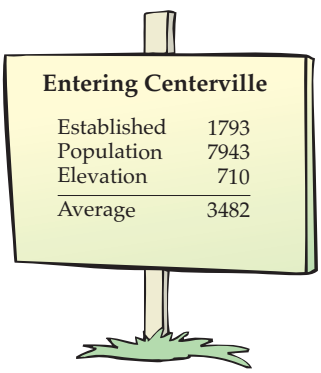
		Class				
		First	Second	Third	Crew	Total
Survival	Alive	203	118	178	212	711
	Dead	122	167	528	673	1490
	Total	325	285	706	885	2201

In each instance, pay attention to the *Who* implicitly defined by the phrase. Often there is a restriction to a smaller group (all aboard the *Titanic*, those in first class, and those who survived, respectively) before a percentage is found. Your discussion of results must make these differences clear.

- **Don't forget to look at the variables separately, too.** When you make a contingency table or display a conditional distribution, be sure you also examine the marginal distributions. It's important to know how many cases are in each category.

- **Be sure to use enough individuals.** When you consider percentages, take care that they are based on a large enough number of individuals. Take care not to make a report such as this one:

  We found that 66.67% of the rats improved their performance with training. The other rat died.



- **Don't overstate your case.** Independence is an important concept, but it is rare for two variables to be *entirely* independent. We can't conclude that one variable has no effect whatsoever on another. Usually, all we know is that little effect was observed in our study. Other studies of other groups under other circumstances could find different results.

Simpson's Paradox

- **Don't use unfair or silly averages.** Sometimes averages can be misleading. Sometimes they just don't make sense at all. Be careful when averaging different variables that the quantities you're averaging are comparable. The Centerville sign says it all.

  When using averages of proportions across several different groups, it's important to make sure that the groups really are comparable.

It's easy to make up an example showing that averaging across very different values or groups can give absurd results. Here's how that might work: Suppose there are two pilots, Moe and Jill. Moe argues that he's the better pilot of the two, since he managed to land 83% of his last 120 flights on time compared with Jill's 78%. But let's look at the data a little more closely. Here are the results for each of their last 120 flights, broken down by the time of day they flew:

Table 3.10
On-time flights by *Time of Day* and *Pilot*. Look at the percentages within each *Time of Day* category. Who has a better on-time record during the day? At night? Who is better overall?

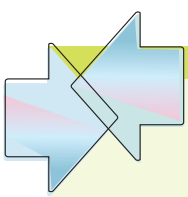
		Time of Day		
		Day	**Night**	**Overall**
Pilot	**Moe**	90 out of 100 90%	10 out of 20 50%	100 out of 120 83%
	Jill	19 out of 20 95%	75 out of 100 75%	94 out of 120 78%

Look at the daytime and nighttime flights separately. For day flights, Jill had a 95% on-time rate and Moe only a 90% rate. At night, Jill was on time 75% of the time and Moe only 50%. So Moe is better "overall," but Jill is better both during the day and at night. How can this be?

What's going on here is a problem known as **Simpson's paradox,** named for the statistician who discovered it in the 1960s. It comes up rarely in real life, but there have been several well-publicized cases. As we can see from the pilot example, the problem is *unfair averaging* over different groups. Jill has mostly night flights, which are more difficult, so her *overall average* is heavily influenced by her nighttime average. Moe, on the other hand, benefits from flying mostly during the day, with its higher on-time percentage. With their very different patterns of flying conditions, taking an overall average is misleading. It's not a fair comparison.

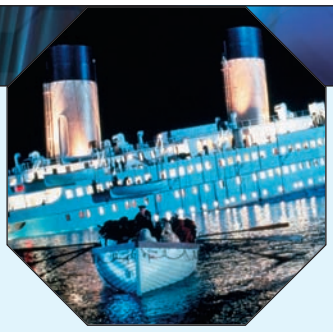
The moral of Simpson's paradox is to be careful when you average across different levels of a second variable. It's always better to compare percentages or other averages *within* each level of the other variable. The overall average may be misleading.

One famous example of Simpson's paradox arose during an investigation of admission rates for men and women at the University of California at Berkeley's graduate schools. As reported in an article in *Science,* about 45% of male applicants were admitted, but only about 30% of female applicants got in. It looked like a clear case of discrimination. However, when the data were broken down by school (Engineering, Law, Medicine, etc.), it turned out that, within each school, the women were admitted at nearly the same or, in some cases, much *higher* rates than the men. How could this be? Women applied in large numbers to schools with very low admission rates (Law and Medicine, for example, admitted fewer than 10%). Men tended to apply to Engineering and Science. Those schools have admission rates above 50%. When the *average* was taken, the women had a much lower *overall* rate, but the average didn't really make sense.



CONNECTIONS

All of the methods of this chapter work with *categorical variables.* You must know the *Who* of the data to know who is counted in each category and the *What* of the variable to know where the categories come from.



WHAT HAVE WE LEARNED?

We've learned that we can summarize categorical data by counting the number of cases in each category, sometimes expressing the resulting distribution as percents. We can display the distribution in a bar chart or a pie chart. When we want to see how two categorical variables are related, we put the counts (and/or percentages) in a two-way table called a contingency table.

- We look at the marginal distribution of each variable (found in the margins of the table).
- We also look at the conditional distribution of a variable within each category of the other variable.
- We can display these conditional and marginal distributions by using bar charts or pie charts.
- If the conditional distributions of one variable are (roughly) the same for every category of the other, the variables are independent.

Terms

Frequency table (Relative frequency table)
21. A frequency table lists the categories in a categorical variable and gives the count (or percentage of observations for each category.

Distribution
22. The distribution of a variable gives
- the possible values of the variable and
- the relative frequency of each value.

Area principle
22. In a statistical display, each data value should be represented by the same amount of area.

Bar chart (Relative frequency bar chart)
22. Bar charts show a bar whose area represents the count (or percentage) of observations for each category of a categorical variable.

Pie chart
23. Pie charts show how a "whole" divides into categories by showing a wedge of a circle whose area corresponds to the proportion in each category.

Categorical data condition
24. The methods in this chapter are appropriate for displaying and describing categorical data. Be careful not to use them with quantitative data.

Contingency table
24. A contingency table displays counts and, sometimes, percentages of individuals falling into named categories on two or more variables. The table categorizes the individuals on all variables at once to reveal possible patterns in one variable that may be contingent on the category of the other.

Marginal distribution
24. In a contingency table, the distribution of either variable alone is called the marginal distribution. The counts or percentages are the totals found in the margins (last row or column) of the table.

Conditional distribution
26. The distribution of a variable restricting the *Who* to consider only a smaller group of individuals is called a conditional distribution.

Independence
29. Variables are said to be independent if the conditional distribution of one variable is the same for each category of the other. We'll show how to check for independence in a later chapter.

Segmented bar chart
30. A segmented bar chart displays the conditional distribution of a categorical variable within each category of another variable.

Simpson's paradox
34. When averages are taken across different groups, they can appear to contradict the overall averages. This is known as "Simpson's paradox."

Skills

THINK
- Be able to recognize when a variable is categorical and choose an appropriate display for it.
- Understand how to examine the association between categorical variables by comparing conditional and marginal percentages.

SHOW
- Be able to summarize the distribution of a categorical variable with a frequency table.
- Be able to display the distribution of a categorical variable with a bar chart or pie chart.
- Know how to make and examine a contingency table.

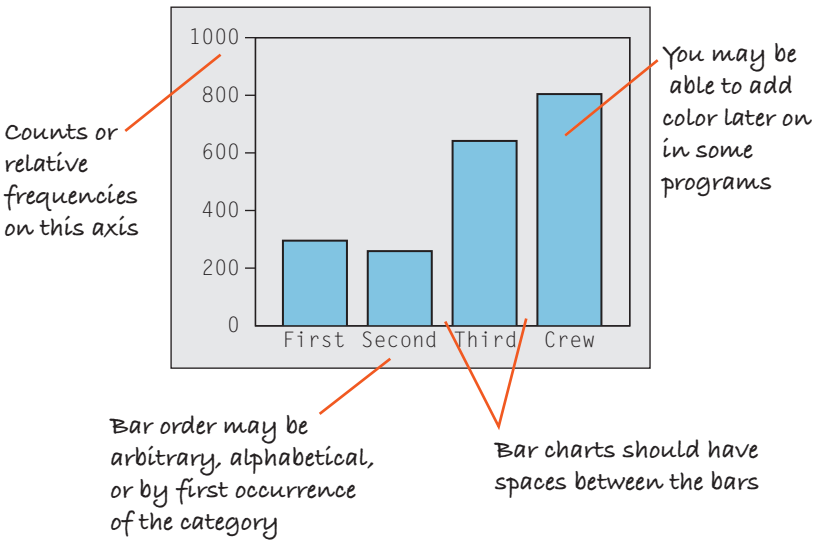
- Know how to make and examine displays of the conditional distributions of one variable for two or more groups.

TELL

- Be able to describe the distribution of a categorical variable in terms of its possible values and relative frequencies.
- Know how to describe any anomalies or extraordinary features revealed by the display of a variable.
- Be able to describe and discuss patterns found in a contingency table and associated displays of conditional distributions.

DISPLAYING CATEGORICAL DATA ON THE COMPUTER

Although every package makes a slightly different bar chart, they all have similar features:



Sometimes the count or a percentage is printed above or on top of each bar to give some additional information. You may find that your statistics package sorts category names in annoying orders by default. For example, many packages sort categories alphabetically or by the order the categories are seen in the data set. Often, neither of these is the best choice.

EXERCISES

1. **Graphs in the news.** Find a bar graph of categorical data from a newspaper, a magazine, or the Internet.
   a) Is the graph clearly labeled?
   b) Does it violate the area principle?
   c) Does the accompanying article tell the W's of the variable?
   d) Do you think the article correctly interprets the data? Explain.

2. **Graphs in the news II.** Find a pie chart of categorical data from a newspaper, a magazine, or the Internet.
   a) Is the graph clearly labeled?
   b) Does it violate the area principle?
   c) Does the accompanying article tell the W's of the variable?
   d) Do you think the article correctly interprets the data? Explain.

CHAPTER

4

Displaying and Summarizing Quantitative Data

Tsunamis are potentially destructive waves that can occur when the sea floor is suddenly and abruptly deformed. They are most often caused by earthquakes beneath the sea that shift the earth's crust, displacing a large mass of water.

The tsunami of December 26, 2004, with epicenter off the west coast of Sumatra, was caused by an earthquake of magnitude 9.0 on the Richter scale. It killed an estimated 297,248 people, making it the most disastrous tsunami on record. But was the earthquake that caused it truly extraordinary, or did it just happen at an unlucky place and time? The U.S. National Geophysical Data Center[1] has information on more than 2400 tsunamis dating back to 2000 B.C.E., and we have estimates of the magnitude of the underlying earthquake for 1240 of them. What can we learn from these data?

Histograms

WHO	1240 earthquakes known to have caused tsunamis for which we have data or good estimates
WHAT	Magnitude (Richter scale [2]), depth (m), date, location, and other variables
WHEN	From 2000 B.C.E. to the present
WHERE	All over the earth

Let's start with a picture. For categorical variables, it is easy to draw the distribution because each category is a natural "pile." But for quantitative variables, there's no obvious way to choose piles. So, usually, we slice up all the possible values into equal-width bins. We then count the number of cases that fall into each bin. The bins, together with these counts, give the **distribution** of the quantitative variable and provide the building blocks for the histogram. By representing the counts as bars and plotting them against the bin values, the **histogram** displays the distribution at a glance.

[1] www.ngdc.noaa.gov

[2] Technically, Richter scale values are in units of log dyne-cm. But the Richter scale is so common now that usually the units are assumed. The U.S. Geological Survey gives the background details of Richter scale measurements on its Web site www.usgs.gov/.

For example, here are the *Magnitudes* (on the Richter scale) of the 1240 earthquakes in the NGDC data:

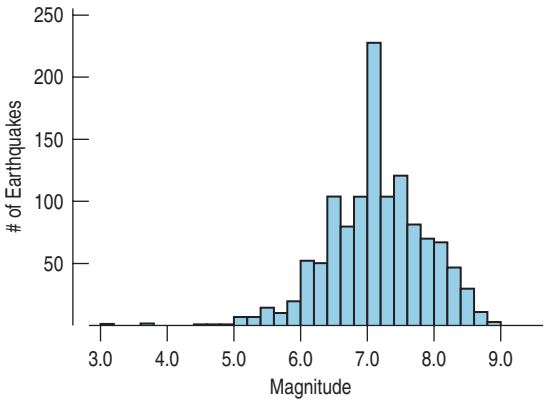


FIGURE 4.1
A histogram of earthquake magnitudes shows the number of earthquakes with magnitudes (in Richter scale units) in each bin.

One surprising feature of the earthquake magnitudes is the spike around magnitude 7.0. Only one other bin holds even half that many earthquakes. These values include historical data for which the magnitudes were estimated by experts and not measured by modern seismographs. Perhaps the experts thought 7 was a typical and reasonable value for a tsunami-causing earthquake when they lacked detailed information. That would explain the overabundance of magnitudes right at 7.0 rather than spread out near that value.

Like a bar chart, a histogram plots the bin counts as the heights of bars. In this histogram of earthquake magnitudes, each bin has a width of 0.2, so, for example, the height of the tallest bar says that there were about 230 earthquakes with magnitudes between 7.0 and 7.2. In this way, the histogram displays the entire distribution of earthquake magnitudes.

Does the distribution look as you expected? 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.

From the histogram, we can see that these earthquakes typically have magnitudes around 7. Most are between 5.5 and 8.5, and some are as small as 3 and as big as 9. Now we can answer the question about the Sumatra tsunami. With a value of 9.0 it's clear that the earthquake that caused it was an extraordinarily powerful earthquake—one of the largest on record.[3]

The bar charts of categorical variables we saw in Chapter 3 had spaces between the bars to separate the counts of different categories. But in a histogram, the bins slice up *all the values* of the quantitative variable, so any spaces in a histogram are actual **gaps** in the data, indicating a region where there are no values.

Sometimes it is useful to make a **relative frequency histogram,** replacing the counts on the vertical axis with the *percentage* of the total number of cases falling in each bin. Of course, the shape of the histogram is exactly the same; only the vertical scale is different.

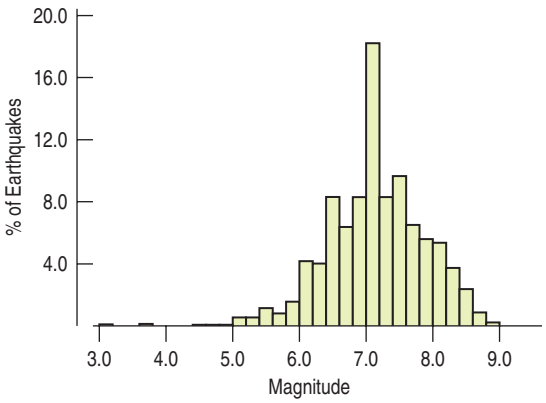
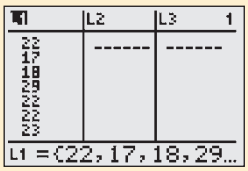


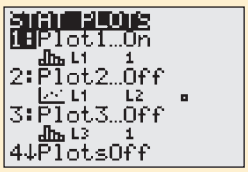
FIGURE 4.2
A relative frequency histogram looks just like a frequency histogram except for the labels on the y-axis, which now show the percentage of earthquakes in each bin.

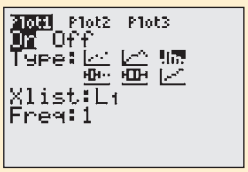
[3] Some experts now estimate the magnitude at between 9.1 and 9.3.

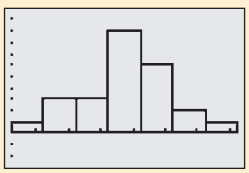
TI Tips

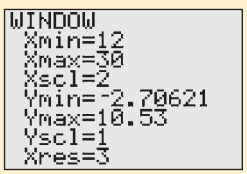
Making a histogram

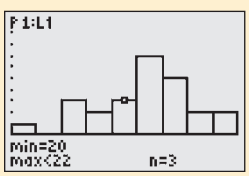


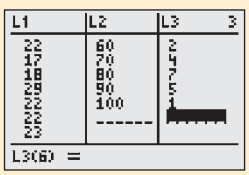


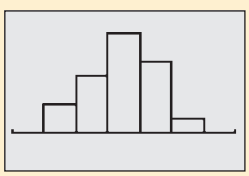












Your calculator can create histograms. First you need some data. For an agility test, fourth-grade children jump from side to side across a set of parallel lines, counting the number of lines they clear in 30 seconds. Here are their scores:

22, 17, 18, 29, 22, 22, 23, 24, 23, 17, 21, 25, 20
12, 19, 28, 24, 22, 21, 25, 26, 25, 16, 27, 22

Enter these data into L1.

Now set up the calculator's plot:

- Go to 2nd STATPLOT, choose Plot1, then ENTER.
- In the Plot1 screen choose On, select the little histogram icon, then specify Xlist:L1 and Freq:1.
- Be sure to turn off any other graphs the calculator may be set up for. Just hit the Y= button, and deactivate any functions seen there.

All set? To create your preliminary plot go to ZOOM, select 9:ZoomStat, and then ENTER.

You now see the calculator's initial attempt to create a histogram of these data. Not bad. We can see that the distribution is roughly symmetric. But it's hard to tell exactly what this histogram shows, right? Let's fix it up a bit.

- Under WINDOW, let's reset the bins to convenient, sensible values. Try Xmin=12, Xmax=30 and Xscl=2. That specifies the range of values along the *x*-axis and makes each bar span two lines.
- Hit GRAPH (*not* ZoomStat—this time we want control of the scale!).

There. We still see rough symmetry, but also see that one of the scores was much lower than the others. Note that you can now find out exactly what the bars indicate by activating TRACE and then moving across the histogram using the arrow keys. For each bar the calculator will indicate the interval of values and the number of data values in that bin. We see that 3 kids had agility scores of 20 or 21.

Play around with the WINDOW settings. A different Ymax will make the bars appear shorter or taller. What happens if you set the bar width (Xscl) smaller? Or larger? You don't want to lump lots of values into just a few bins or make so many bins that the overall shape of the histogram is not clear. Choosing the best bar width takes practice.

Finally, suppose the data are given as a frequency table. Consider a set of test scores, with two grades in the 60s, four in the 70s, seven in the 80s, five in the 90s, and one 100. Enter the group cutoffs 60, 70, 80, 90, 100 in L2 and the corresponding frequencies 2, 4, 7, 5, 1 in L3. When you set up the histogram STATPLOT, specify Xlist:L2 and Freq:L3. Can you specify the WINDOW settings to make this histogram look the way you want it? (By the way, if you get a DIM MISMATCH error, it means you can't count. Look at L2 and L3; you'll see the two lists don't have the same number of entries. Fix the problem by correcting the data you entered.)

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. Here's a histogram of the pulse rates of 24 women, taken by a researcher at a health clinic:

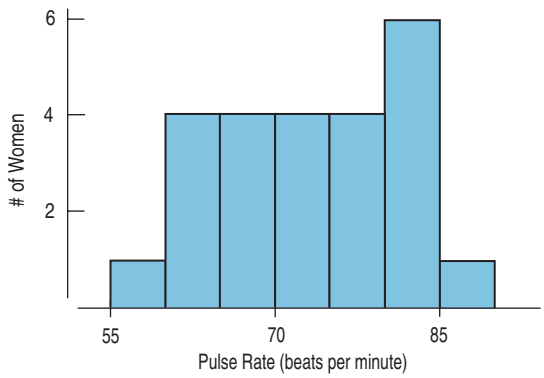


FIGURE 4.3
The pulse rates of 24 women at a health clinic

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, but we prefer Tukey's original name for it.

The story seems pretty clear. We can see the entire span of the data and can easily see what a typical pulse rate might be. But is that all there is to these data?

A **stem-and-leaf display** is like a histogram, but it shows the individual values. It's also easier to make by hand. Here's a stem-and-leaf display of the same data:

8	8
8	000044
7	6666
7	2222
6	8888
6	0444
5	6

Pulse Rate
(8|8 means 88 beats/min)

A S ***Activity:* Stem-and-Leaf Displays.** As you might expect of something called "stem-and leaf," these displays grow as you consider each data value.

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? Well, it's backwards because now the higher values are on the left, but other than that, it has the same shape.[4]

What does the line at the top of the display that says 8 | 8 mean? It stands for a pulse of 88 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 8 | 000044. That shows that there were four pulse rates of 80 and two of 84 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

[4] You could make the stem-and-leaf with the higher values on the bottom. Usually, though, higher on the top makes sense.

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?

How do stem-and-leaf displays work? Stem-and-leaf displays work like histograms, but they show more information. They use part of the number itself (called the stem) to name the bins. To make the "bars," they use the next digit of the number. For example, if we had a test score of 83, we could write it 8|3, where 8 serves as the stem and 3 as the leaf. Then, to display the scores 83, 76, and 88 together, we would write

```
8 | 38
7 | 6
```

For the pulse data, we have

```
8 | 0000448
7 | 22226666
6 | 04448888
5 | 6
Pulse Rate
(5|6 means 56 beats/min)
```

This display is OK, but a little crowded. A histogram might split each line into two bars. With a stem-and-leaf, we can do the same by putting the leaves 0–4 on one line and 5–9 on another, as we saw above:

```
8 | 8
8 | 000044
7 | 6666
7 | 2222
6 | 8888
6 | 0444
5 | 6
Pulse Rate
(8|8 means 88 beats/min)
```

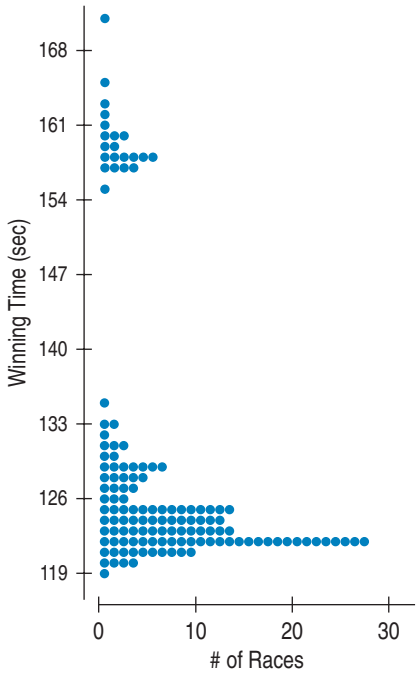
For numbers with three or more digits, you'll often decide to truncate (or round) the number to two places, using the first digit as the stem and the second as the leaf. So, if you had 432, 540, 571, and 638, you might display them as shown below with an indication that 6|3 means 630–639.

```
6 | 3
5 | 47
4 | 3
```

When you make a stem-and-leaf by hand, make sure to give each digit the same width, in order to preserve the area principle. (That can lead to some fat 1's and thin 8's—but it makes the display honest.)

Dotplots

Activity: Dotplots. Click on points to see their values and even drag them around.



A **dotplot** is a simple display. It just 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 (especially if you forget how to write the digits from 0 to 9). Here's 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 2008 Derby.

Dotplots show basic facts about the distribution. We can find the slowest and quickest races by finding times for the topmost and bottommost dots. It's also 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.

Some dotplots stretch out horizontally, with the counts on the vertical axis, like a histogram. Others, such as the one shown here, run vertically, like a stem-and-leaf display. Some dotplots place points next to each other when they would otherwise overlap. Others just place them on top of one another. Newspapers sometimes offer dotplots with the dots made up of little pictures.

FIGURE 4.4
A dotplot of Kentucky Derby winning times plots each race as its own dot, showing the bimodal distribution.

Think Before You Draw, Again

Suddenly, we face a lot more options when it's time to invoke our first rule of data analysis and make a picture. You'll need to *Think* carefully to decide which type of graph to make. In the previous chapter you learned to check the Categorical Data Condition before making a pie chart or a bar chart. Now, before making a stem-and-leaf display, a histogram, or a dotplot, you need to check the

> **Quantitative Data Condition:** The data are values of a quantitative variable whose units are known.

Although a bar chart and a histogram may look somewhat similar, they're not the same display. You can't display categorical data in a histogram or quantitative data in a bar chart. Always check the condition that confirms what type of data you have before proceeding with your display.

Step back from a histogram or stem-and-leaf display. What can you say about the distribution? When you describe a distribution, you should always tell about three things: its **shape, center,** and **spread.**

The Shape of a Distribution

1. *Does the histogram have a single, central hump or several separated humps?* These humps are called **modes**.[5] The earthquake magnitudes have a single mode

[5] 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.

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. What is the mode of the Kentucky Derby times? Well, seven races were timed at 122.2 seconds—more than any other race time. Should that be the mode? Probably not. For quantitative data, it makes more sense to use the term "mode" in the more general sense of the peak of the histogram rather than as a single summary value. In this sense, 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.

at just about 7. A histogram with one peak, such as the earthquake magnitudes, is dubbed **unimodal;** histograms with two peaks are **bimodal,** and those with three or more are called **multimodal**.[6] For example, here's a bimodal histogram.

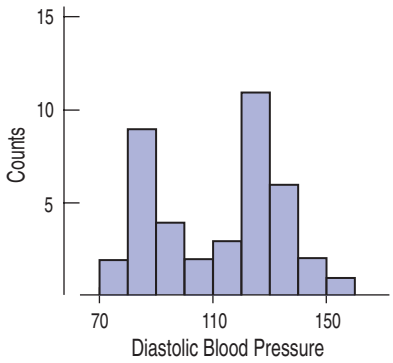


FIGURE 4.5
A bimodal histogram has two apparent peaks.

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

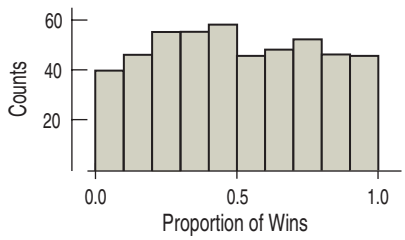


FIGURE 4.6
In a uniform histogram, the bars are all about the same height. The histogram doesn't appear to have a mode.

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.

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?

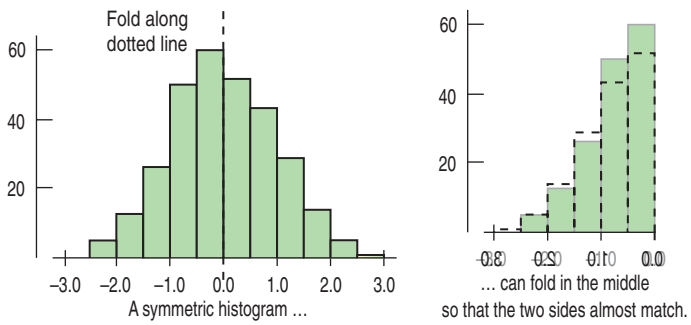


FIGURE 4.7

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.

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

A S ***Activity:*** **Attributes of Distribution Shape.** This activity and the others on this page show off aspects of distribution shape through animation and example, then let you make and interpret histograms with your statistics package.

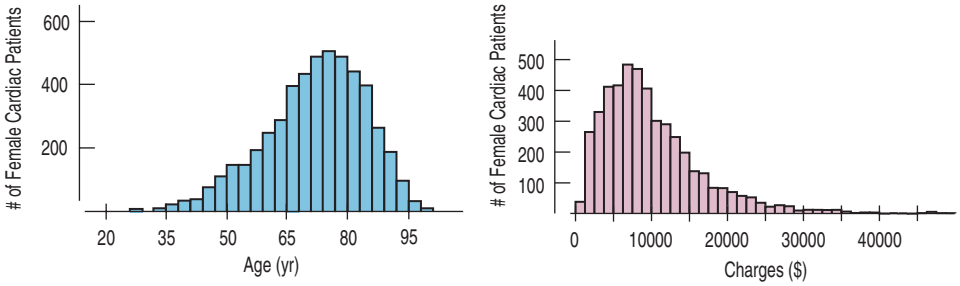
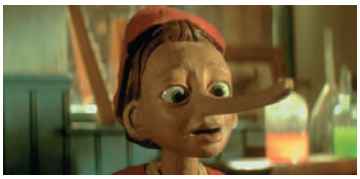


FIGURE 4.8

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 tell us something interesting or exciting about the data. You should always mention any stragglers, or **outliers,** that stand off away from the body of the distribution. If you're collecting data on nose lengths and Pinocchio is in the group, you'd probably notice him, 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. But don't throw it away without comment. Treat it specially and discuss it when you tell about your data. Or find the error and fix it if you can. Be sure to look for outliers. Always.

   In the next chapter you'll learn a handy rule of thumb for deciding when a point might be considered an outlier.

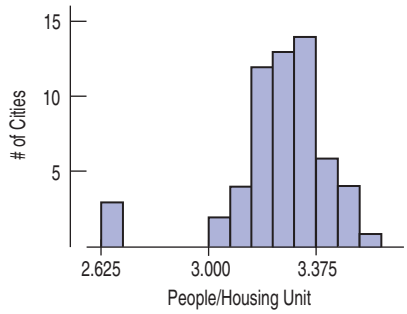


FIGURE 4.9

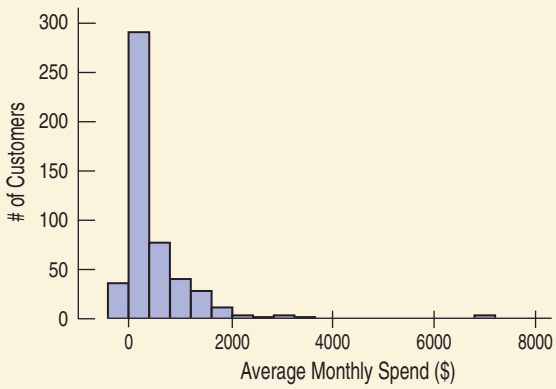
A histogram with outliers. There are three cities in the leftmost bar.

FOR EXAMPLE **Describing histograms**

A credit card company wants to see how much customers in a particular segment of their market use their credit card. They have provided you with data[7] on the amount spent by 500 selected customers during a 3-month period and have asked you to summarize the expenditures. Of course, you begin by making a histogram.

Question: Describe the shape of this distribution.

The distribution of expenditures is unimodal and skewed to the high end. There is an extraordinarily large value at about $7000, and some of the expenditures are negative.



[7] These data are real, but cannot be further identified for obvious privacy reasons.

Are there any gaps in the distribution? The Kentucky Derby data that we saw in the dotplot on page 49 has a large gap between two groups of times, one near 120 seconds and one near 160. Gaps help us see multiple modes and encourage us to notice when the data may come from different sources or contain more than one group.

Toto, I've a feeling we're not in math class anymore . . . When Dorothy and her dog Toto land in Oz, everything is more vivid and colorful, but also more dangerous and exciting. Dorothy has new choices to make. She can't always rely on the old definitions, and the yellow brick road has many branches. You may be coming to a similar realization about Statistics.

When we summarize data, our goal is usually more than just developing a detailed knowledge of the data we have at hand. Scientists generally don't care about the particular guinea pigs they've treated, but rather about what their reactions say about how animals (and, perhaps, humans) would respond.

When you look at data, you want to know what the data say about the world, so you'd like to know whether the patterns you see in histograms and summary statistics generalize to other individuals and situations. You'll want to calculate summary statistics accurately, but then you'll also want to think about what they may say beyond just describing the data. And your knowledge about the world matters when you think about the overall meaning of your analysis.

It may surprise you that many of the most important concepts in Statistics are not defined as precisely as most concepts in mathematics. That's done on purpose, to leave room for judgment.

Because we want to see broader patterns rather than focus on the details of the data set we're looking at, we deliberately leave some statistical concepts a bit vague. Whether a histogram is symmetric or skewed, whether it has one or more modes, whether a point is far enough from the rest of the data to be considered an outlier—these are all somewhat vague concepts. And they all require judgment. You may be used to finding a single correct and precise answer, but in Statistics, there may be more than one interpretation. That may make you a little uncomfortable at first, but soon you'll see that this room for judgment brings you enormous power and responsibility. It means that using your own knowledge and judgment and supporting your findings with statistical evidence and justifications entitles you to your own opinions about what you see.

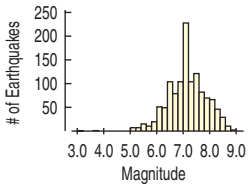
JUST CHECKING

It's often a good idea to think about what the distribution of a data set might look like before we collect the data. What do you think the distribution of each of the following data sets will look like? Be sure to discuss its shape. Where do you think the center might be? How spread out do you think the values will be?

1. Number of miles run by Saturday morning joggers at a park.
2. Hours spent by U.S. adults watching football on Thanksgiving Day.
3. Amount of winnings of all people playing a particular state's lottery last week.
4. Ages of the faculty members at your school.
5. Last digit of phone numbers on your campus.

The Center of the Distribution: The Median

Let's return to the tsunami earthquakes. But this time, let's look at just 25 years of data: 176 earthquakes that occurred from 1981 through 2005. These should be more accurately measured than prehistoric quakes because seismographs were in wide use. Try to put your finger on the histogram at the value you think is



typical. (Read the value from the horizontal axis and remember it.) When we think of a typical value, we usually look for the **center** of the distribution. Where do you think the center of this distribution is? For a unimodal, symmetric distribution such as these earthquake data, it's easy. We'd all agree on the center of symmetry, where we would fold the histogram to match the two sides. But when the distribution is skewed or possibly multimodal, it's not immediately clear what we even mean by the center.

One reasonable choice of typical value is the value that is literally in the middle, with half the values below it and half above it.

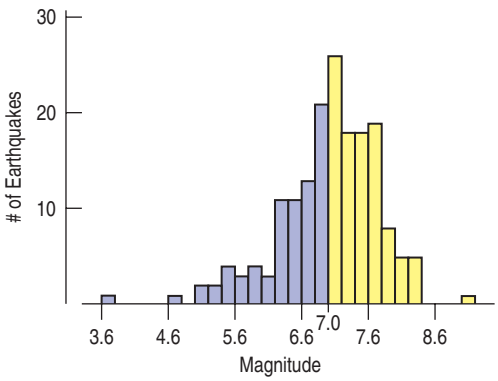


FIGURE 4.10 ***Tsunami-causing earthquakes (1981–2005)***

The median splits the histogram into two halves of equal area.

Histograms follow the area principle, and each half of the data has about 88 earthquakes, so each colored region has the same area in the display. The middle value that divides the histogram into two equal areas is called the **median.**

The median has the same units as the data. Be sure to include the units whenever you discuss the median.

For the recent tsunamis, there are 176 earthquakes, so the median is found at the $(176 + 1)/2 = 88.5$th place in the sorted data. That ".5" just says to average the two values on either side: the 88th and the 89th. The median earthquake magnitude is 7.0.

NOTATION ALERT:

We always use n to indicate the number of values. Some people even say, "How big is the n?" when they mean the number of data values.

How do medians work? Finding the median of a batch of n numbers is easy as long as you remember to order the values first. If n is odd, the median is the middle value. Counting in from the ends, we find this value in the $\frac{n+1}{2}$ position.

When n is even, there are two middle values. So, in this case, the median is the average of the two values in positions $\frac{n}{2}$ and $\frac{n}{2} + 1$.

Here are two examples:

Suppose the batch has these values: 14.1, 3.2, 25.3, 2.8, −17.5, 13.9, 45.8.
First we order the values: −17.5, 2.8, 3.2, 13.9, 14.1, 25.3, 45.8.
Since there are 7 values, the median is the $(7 + 1)/2 = 4$th value, counting from the top or bottom: 13.9. Notice that 3 values are lower, 3 higher.

Suppose we had the same batch with another value at 35.7. Then the ordered values are −17.5, 2.8, 3.2, 13.9, 14.1, 25.3, 35.7, 45.8.
The median is the average of the 8/2 or 4th, and the $(8/2) + 1$, or 5th, values. So the median is $(13.9 + 14.1)/2 = 14.0$. Four data values are lower, and four higher.

The median is one way to find the center of the data. But there are many others. We'll look at an even more important measure later in this chapter.

Knowing the median, we could say that a typical tsunami-causing earthquake, worldwide, was about 7.0 on the Richter scale. How much does that really say? How well does the median describe the data? After all, not every earthquake has a Richter scale value of 7.0. Whenever we find the center of data, the next step is always to ask how well it actually summarizes the data.

Spread: Home on the Range

Statistics pays close attention to what we *don't* know as well as what we do know. Understanding how spread out the data are is a first step in understanding what a summary *cannot* tell us about the data. It's the beginning of telling us what we don't know.

If every earthquake that caused a tsunami registered 7.0 on the Richter scale, then knowing the median would tell us everything about the distribution of earthquake magnitudes. The more the data vary, however, the less the median alone can tell us. So we need to measure how much the data values vary around the center. In other words, how spread out are they? When we describe a distribution numerically, we always report a measure of its **spread** along with its center.

How should we measure the spread? We could simply look at the extent of the data. How far apart are the two extremes? The **range** of the data is defined as the *difference* between the maximum and minimum values:

$$Range = max - min.$$

Notice that the range is a *single number, not* an interval of values, as you might think from its use in common speech. The maximum magnitude of these earthquakes is 9.0 and the minimum is 3.7, so the *range* is $9.0 - 3.7 = 5.3$.

The range has the disadvantage that a single extreme value can make it very large, giving a value that doesn't really represent the data overall.

Spread: The Interquartile Range

A better way to describe the spread of a variable might be to ignore the extremes and concentrate on the middle of the data. We could, for example, find the range of just the middle half of the data. What do we mean by the middle half? Divide the data in half at the median. Now divide both halves in half again, cutting the data into four quarters. We call these new dividing points **quartiles**. One quarter of the data lies below the **lower quartile,** and one quarter of the data lies above the **upper quartile,** so half the data lies between them. The quartiles border the middle half of the data.

How do quartiles work? A simple way to find the quartiles is to start by splitting the batch into two halves at the median. (When n is odd, some statisticians include the median in both halves; others omit it.) The lower quartile is the median of the lower half, and the upper quartile is the median of the upper half.

Here are our two examples again.

The ordered values of the first batch were −17.5, 2.8, 3.2, 13.9, 14.1, 25.3, and 45.8, with a median of 13.9. Excluding the median, the two halves of the list are −17.5, 2.8, 3.2 and 14.1, 25.3, 45.8.

Each half has 3 values, so the median of each is the middle one. The lower quartile is 2.8, and the upper quartile is 25.3.

The second batch of data had the ordered values −17.5, 2.8, 3.2, 13.9, 14.1, 25.3, 35.7, and 45.8.

Here n is even, so the two halves of 4 values are −17.5, 2.8, 3.2, 13.9 and 14.1, 25.3, 35.7, 45.8.

Now the lower quartile is $(2.8 + 3.2)/2 = 3.0$, and the upper quartile is $(25.3 + 35.7)/2 = 30.5$.

The difference between the quartiles tells us how much territory the middle half of the data covers and is called the **interquartile range.** It's commonly abbreviated IQR (and pronounced "eye-cue-are," not "ikker"):

$$IQR = \text{upper quartile} - \text{lower quartile}.$$

For the earthquakes, there are 88 values below the median and 88 values above the median. The midpoint of the lower half is the average of the 44th and 45th values in the ordered data; that turns out to be 6.6. In the upper half we average the 132nd and 133rd values, finding a magnitude of 7.6 as the third quartile. The *difference* between the quartiles gives the IQR:

$$IQR = 7.6 - 6.6 = 1.0.$$

Now we know that the middle half of the earthquake magnitudes extends across a (interquartile) range of 1.0 Richter scale units. This seems like a reasonable summary of the spread of the distribution, as we can see from this histogram:

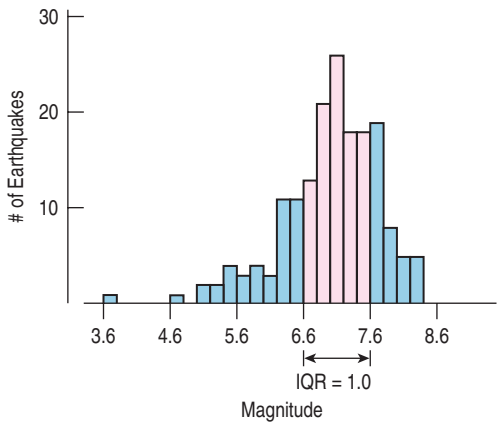


FIGURE 4.11
The quartiles bound the middle 50% of the values of the distribution. This gives a visual indication of the spread of the data. Here we see that the IQR is 1.0 Richter scale units.

The IQR is almost always a reasonable summary of the spread of a distribution. Even if the distribution itself is skewed or has some outliers, the IQR should provide useful information. The one exception is when the data are strongly bimodal. For example, remember the dotplot of winning times in the Kentucky Derby (page 49)? Because the race distance was changed, we really have data on two different races, and they shouldn't be summarized together.

So, what is a quartile anyway? Finding the quartiles sounds easy, but surprisingly, the quartiles are not well-defined. It's not always clear how to find a value such that exactly one quarter of the data lies above or below that value. We offered a simple rule for Finding Quartiles in the box on page 54: Find the median of each half of the data split by the median. When n is odd, we (and your TI calculator) omit the median from each of the halves. Some other texts include the median in both halves before finding the quartiles. Both methods are commonly used. If you are willing to do a bit more calculating, there are several other methods that locate a quartile somewhere between adjacent data values. We know of at least six different rules for finding quartiles. Remarkably, each one is in use in some software package or calculator.

So don't worry too much about getting the "exact" value for a quartile. All of the methods agree pretty closely when the data set is large. When the data set is small, different rules will disagree more, but in that case there's little need to summarize the data anyway.

Remember, Statistics is about understanding the world, not about calculating the right number. The "answer" to a statistical question is a sentence about the issue raised in the question.

The lower and upper quartiles are also known as the 25th and 75th **percentiles** of the data, respectively, since the lower quartile falls above 25% of the data and the upper quartile falls above 75% of the data. If we count this way, the median is the 50th percentile. We could, of course, define and calculate any percentile that we want. For example, the 10th percentile would be the number that falls above the lowest 10% of the data values.

5-Number Summary

NOTATION ALERT:

We always use Q1 to label the lower (25%) quartile and Q3 to label the upper (75%) quartile. We skip the number 2 because the median would, by this system, naturally be labeled Q2—but we don't usually call it that.

The **5-number summary** of a distribution reports its median, quartiles, and extremes (maximum and minimum). The 5-number summary for the recent tsunami earthquake *Magnitudes* looks like this:

Max	9.0
Q3	7.6
Median	7.0
Q1	6.6
Min	3.7

It's good idea to report the number of data values and the identity of the cases (the *Who*). Here there are 176 earthquakes.

The 5-number summary provides a good overview of the distribution of magnitudes of these tsunami-causing earthquakes. For a start, we can see that the median magnitude is 7.0. Because the IQR is only $7.6 - 6.6 = 1$, we see that many quakes are close to the median magnitude. Indeed, the quartiles show us that the middle half of these earthquakes had magnitudes between 6.6 and 7.6. One quarter of the earthquakes had magnitudes above 7.6, although one tsunami was caused by a quake measuring only 3.7 on the Richter scale.

STEP-BY-STEP EXAMPLE **Shape, Center, and Spread: Flight Cancellations**

The U.S. Bureau of Transportation Statistics (www.bts.gov) reports data on airline flights. Let's look at data giving the percentage of flights cancelled each month between 1995 and 2005.

Question: How often are flights cancelled?

WHO	Months
WHAT	Percentage of flights cancelled at U.S. airports
WHEN	1995–2005
WHERE	United States

THINK

Variable: Identify the *variable*, and decide how you wish to display it.

To identify a variable, report the W's.

I want to learn about the monthly percentage of flight cancellations at U.S airports.

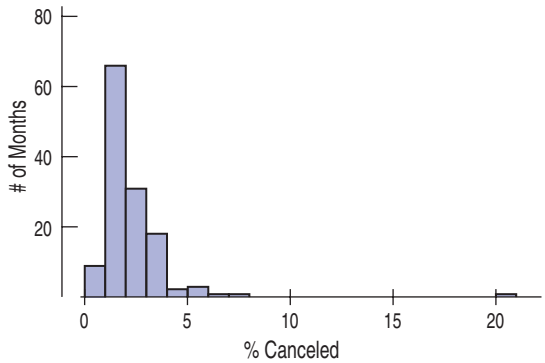
I have data from the U.S. Bureau of Transportation Statistics giving the percentage of flights cancelled at U.S. airports each month between 1995 and 2005.

Select an appropriate display based on the nature of the data and what you want to know.

✓ **Quantitative Data Condition:** Percentages are quantitative. A histogram and numerical summaries would be appropriate.

SHOW

Mechanics: We usually make histograms with a computer or graphing calculator.



The histogram shows a distribution skewed to the high end and one extreme outlier, a month in which more than 20% of flights were cancelled.

REALITY CHECK

It's always a good idea to think about what you expect to see so that you can check whether the histogram looks like what you expected.

In most months, fewer than 5% of flights are cancelled and usually only about 2% or 3%. That seems reasonable.

With 132 cases, we probably have more data than you'd choose to work with by hand. The results given here are from technology.

Count	132
Max	20.240
Q3	2.615
Median	1.755
Q1	1.445
Min	0.770
IQR	1.170

Interpretation: Describe the shape, center, and spread of the distribution. Report on the symmetry, number of modes, and any gaps or outliers. You should also mention any concerns you may have about the data.

The distribution of cancellations is skewed to the right, and this makes sense: The values can't fall below 0%, but can increase almost arbitrarily due to bad weather or other events.

The median is 1.76% and the IQR is 1.17%. The low IQR indicates that in most months the cancellation rate is close to the median. In fact, it's between 1.4% and 2.6% in the middle 50% of all months, and in only 1/4 of the months were more than 2.6% of flights cancelled.

There is one extraordinary value: 20.2%. Looking it up, I find that the extraordinary month was September 2001. The attacks of September 11 shut down air travel for several days, accounting for this outlier.

Summarizing Symmetric Distributions: The Mean

NOTATION ALERT:

In Algebra you used letters to represent values in a problem, but it didn't matter what letter you picked. You could call the width of a rectangle X or you could call it w (or *Fred*, for that matter). But in Statistics, the notation is part of the vocabulary. For example, in Statistics n is always the number of data values. Always.

We have already begun to point out such special notation conventions: n, Q1, and Q3. Think of them as part of the terminology you need to learn in this course.

Here's another one: Whenever we put a bar over a symbol, it means "find the mean."

Medians do a good job of summarizing the center of a distribution, even when the shape is skewed or when there is an outlier, as with the flight cancellations. But when we have symmetric data, there's another alternative. You probably already know how to average values. In fact, to find the median when n is even, we said you should average the two middle values, and you didn't even flinch.

The earthquake magnitudes are pretty close to symmetric, so we can also summarize their center with a mean. The mean tsunami earthquake magnitude is 6.96—about what we might expect from the histogram. You already know how to average values, but this is a good place to introduce notation that we'll use throughout the book. We use the Greek capital letter sigma, Σ, to mean "sum" (sigma is "S" in Greek), and we'll write:

$$\bar{y} = \frac{Total}{n} = \frac{\sum y}{n}.$$

The formula says to add up all the values of the variable and divide that sum by the number of data values, n—just as you've always done.[8]

Once we've averaged the data, you'd expect the result to be called the *average*, but that would be too easy. Informally, we speak of the "average person" but we don't add up people and divide by the number of people. A median is also a kind of average. To make this distinction, the value we calculated is called the mean, $\bar{y}$, and pronounced "y-bar."

[8] You may also see the variable called x and the equation written $\bar{x} = \frac{Total}{n} = \frac{\sum x}{n}$. Don't let that throw you. You are free to name the variable anything you want, but we'll generally use y for variables like this that we want to summarize, model, or predict. (Later we'll talk about variables that are used to explain, model, or predict y. We'll call them x.)

The **mean** feels like the center because it is the point where the histogram balances:

In everyday language, sometimes "average" *does* mean what we want it to mean. We don't talk about your grade point mean or a baseball player's batting mean or the Dow Jones Industrial mean. So we'll continue to say "average" when that seems most natural. When we do, though, you may assume that what we mean is the mean.

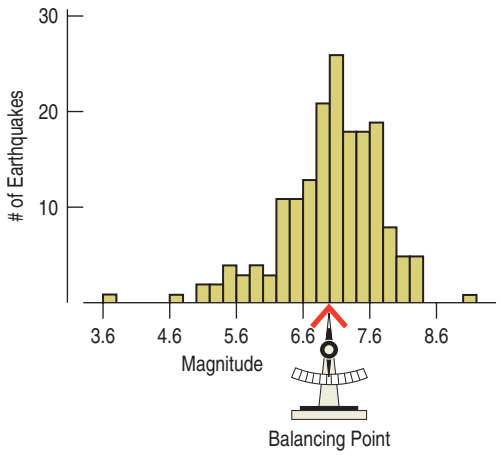


FIGURE 4.12
The mean is located at the balancing point *of the histogram.*

Mean or Median?

Using the center of balance makes sense when the data are symmetric. But data are not always this well behaved. If the distribution is skewed or has outliers, the center is not so well defined and the mean may not be what we want. For example, the mean of the flight cancellations doesn't give a very good idea of the typical percentage of cancellations.

TI-*nspire*

Mean, median, and outliers. Drag data points around to explore how outliers affect the mean and median.

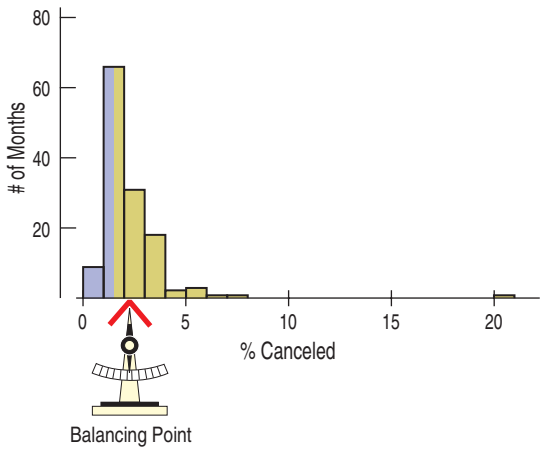


FIGURE 4.13
The median splits the area of the histogram in half at 1.755%. Because the distribution is skewed to the right, the mean (2.28%) is higher *than the median. The points at the right have pulled the mean toward them away from the median.*

A S

***Activity*: The Center of a Distribution.** Compare measures of center by dragging points up and down and seeing the consequences. Another activity shows how to find summaries with your statistics package.

The mean is 2.28%, but nearly 70% of months had cancellation rates below that, so the mean doesn't feel like a good overall summary. Why is the balancing point so high? The large outlying value pulls it to the right. For data like these, the median is a better summary of the center.

Because the median considers only the order of the values, it is **resistant** to values that are extraordinarily large or small; it simply notes that they are one of the "big ones" or the "small ones" and ignores their distance from the center.

For the tsunami earthquake magnitudes, it doesn't seem to make much difference—the mean is 6.96; the median is 7.0. When the data are symmetric, the mean and median will be close, but when the data are skewed, the median is likely to be a better choice. So, why not just use the median? Well, for one, the median can go overboard. It's not just resistant to occasional outliers, but can be unaffected by changes in up to half the data values. By contrast, the mean includes input from

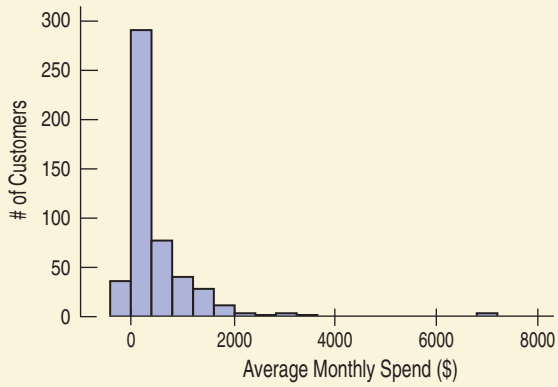
each data value and gives each one equal weight. It's also easier to work with, so when the distribution is unimodal and symmetric, we'll use the mean.

Of course, to choose between mean and median, we'll start by looking at the data. If the histogram is symmetric and there are no outliers, we'll prefer the mean. However, if the histogram is skewed or has outliers, we're usually better off with the median. If you're not sure, report both and discuss why they might differ.

FOR EXAMPLE Describing center

Recap: You want to summarize the expenditures of 500 credit card company customers, and have looked at a histogram.

Question: You have found the mean expenditure to be $478.19 and the median to be $216.28. Which is the more appropriate measure of center, and why?



Because the distribution of expenditures is skewed, the median is the more appropriate measure of center. Unlike the mean, it's not affected by the large outlying value or by the skewness. Half of these credit card customers had average monthly expenditures less than $216.28 and half more.

When to expect skewness Even without making a histogram, we can expect some variables to be skewed. When values of a quantitative variable are bounded on one side but not the other, the distribution may be skewed. For example, incomes and waiting times can't be less than zero, so they are often skewed to the right. Amounts of things (dollars, employees) are often skewed to the right for the same reason. If a test is too easy, the distribution will be skewed to the left because many scores will bump against 100%. And combinations of things are often skewed. In the case of the cancelled flights, flights are more likely to be cancelled in January (due to snowstorms) and in August (thunderstorms). Combining values across months leads to a skewed distribution.

What About Spread? The Standard Deviation

A S ***Activity:* The Spread of a Distribution.** What happens to measures of spread when data values change may not be quite what you expect.

The IQR is always a reasonable summary of spread, but because it uses only the two quartiles of the data, it ignores much of the information about how individual values vary. A more powerful approach uses the **standard deviation,** which takes into account how far *each* value is from the mean. Like the mean, the standard deviation is appropriate only for symmetric data.

One way to think about spread is to examine how far each data value is from the mean. This difference is called a *deviation*. We could just average the deviations, but the positive and negative differences always cancel each other out. So the average deviation is always zero—not very helpful.

To keep them from canceling out, we *square* each deviation. Squaring always gives a positive value, so the sum won't be zero. That's great. Squaring also emphasizes larger differences—a feature that turns out to be both good and bad.

NOTATION ALERT:

s^2 always means the variance of a set of data, and s always denotes the standard deviation.

When we add up these squared deviations and find their average (almost), we call the result the **variance:**

$$s^2 = \frac{\sum(y - \bar{y})^2}{n - 1}.$$

Why almost? It *would* be a mean if we divided the sum by n. Instead, we divide by $n - 1$. Why? The simplest explanation is "to drive you crazy." But there are good technical reasons, some of which we'll see later.

The variance will play an important role later in this book, but it has a problem as a measure of spread. Whatever the units of the original data are, the variance is in *squared* units. We want measures of spread to have the same units as the data. And we probably don't want to talk about squared dollars or mpg^2. So, to get back to the original units, we take the square root of s^2. The result, s, is the **standard deviation.**

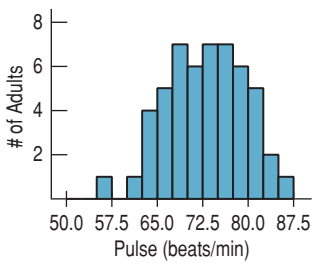
Putting it all together, the standard deviation of the data is found by the following formula:

$$s = \sqrt{\frac{\sum(y - \bar{y})^2}{n - 1}}.$$

You will almost always rely on a calculator or computer to do the calculating.

Understanding what the standard deviation really means will take some time, and we'll revisit the concept in later chapters. For now, have a look at this histogram of resting pulse rates. The distribution is roughly symmetric, so it's okay to choose the mean and standard deviation as our summaries of center and spread. The mean pulse rate is 72.7 beats per minute, and we can see that's a typical heart rate. We also see that some heart rates are higher and some lower—but how much? Well, the standard deviation of 6.5 beats per minute indicates that, on average, we might expect people's heart rates to differ from the mean rate by about 6.5 beats per minute. Looking at the histogram, we can see that 6.5 beats above or below the mean appears to be a typical deviation.

WHO 52 adults
WHAT Resting heart rates
UNITS Beats per minute



How does standard deviation work? To find the standard deviation, start with the mean, $\bar{y}$. Then find the *deviations* by taking $\bar{y}$ from each value: $(y - \bar{y})$ Square each deviation: $(y - \bar{y})^2$.

Now you're nearly home. Just add these up and divide by $n - 1$. That gives you the variance, s^2. To find the standard deviation, s, take the square root. Here we go:

Suppose the batch of values is 14, 13, 20, 22, 18, 19, and 13.

The mean is $\bar{y} = 17$. So the deviations are found by subtracting 17 from each value:

Original Values	Deviations	Squared Deviations
14	$14 - 17 = -3$	$(-3)^2 = 9$
13	$13 - 17 = -4$	$(-4)^2 = 16$
20	$20 - 17 = 3$	9
22	$22 - 17 = 5$	25
18	$18 - 17 = 1$	1
19	$19 - 17 = 2$	4
13	$13 - 17 = -4$	16

Add up the squared deviations: $9 + 16 + 9 + 25 + 1 + 4 + 16 = 80$.

Now divide by $n - 1$: $80/6 = 13.33$.

Finally, take the square root: $s = \sqrt{13.33} = 3.65$

Thinking About Variation

A S ***Activity:* Displaying Spread.** What does the standard deviation look like on a histogram? How about the IQR?

Statistics is about variation, so spread is an important fundamental concept in Statistics. Measures of spread help us to be precise about what we *don't* know. If many data values are scattered far from the center, the IQR and the standard deviation will be large. If the data values are close to the center, then these measures of spread will be small. If all our data values were exactly the same, we'd have no question about summarizing the center, and all measures of spread would be zero—and we wouldn't need Statistics. You might think this would be a big plus, but it would make for a boring world. Fortunately (at least for Statistics), data do vary.

Why do banks favor a single line that feeds several teller windows rather than separate lines for each teller? The average waiting time is the same. But the time you can expect to wait is less variable when there is a single line, and people prefer consistency.

Measures of spread tell how well other summaries describe the data. That's why we always (always!) report a spread along with any summary of the center.

JUST CHECKING

6. The U.S. Census Bureau reports the median family income in its summary of census data. Why do you suppose they use the median instead of the mean? What might be the disadvantages of reporting the mean?

7. You've just bought a new car that claims to get a highway fuel efficiency of 31 miles per gallon. Of course, your mileage will "vary." If you had to guess, would you expect the IQR of gas mileage attained by all cars like yours to be 30 mpg, 3 mpg, or 0.3 mpg? Why?

8. A company selling a new MP3 player advertises that the player has a mean lifetime of 5 years. If you were in charge of quality control at the factory, would you prefer that the standard deviation of lifespans of the players you produce be 2 years or 2 months? Why?

What to *Tell* About a Quantitative Variable

A S ***Activity:* Playing with Summaries.** Here's a Statistics game about summaries that even some experienced statisticians find . . . well, challenging. Your intuition may be better. Give it a try!

What should you *Tell* about a quantitative variable?

- Start by making a histogram or stem-and-leaf display, and discuss the shape of the distribution.
- Next, discuss the center *and* spread.
  - We always pair the median with the IQR and the mean with the standard deviation. It's not useful to report one without the other. Reporting a center without a spread is dangerous. You may think you know more than you do about the distribution. Reporting only the spread leaves us wondering where we are.
  - If the shape is skewed, report the median and IQR. You may want to include the mean and standard deviation as well, but you should point out why the mean and median differ.
  - If the shape is symmetric, report the mean and standard deviation and possibly the median and IQR as well. For unimodal symmetric data, the IQR is usually a bit larger than the standard deviation. If that's not true of your data set, look again to make sure that the distribution isn't skewed and there are no outliers.

TI-*nspire*

Standard deviation, IQR, and outliers. Drag data points around to explore how outliers affect measures of spread.

How "Accurate" Should We Be?

Don't think you should report means and standard deviations to a zillion decimal places; such implied accuracy is really meaningless. Although there is no ironclad rule, statisticians commonly report summary statistics to one or two decimal places more than the original data have.

- Also, discuss any unusual features.
  - If there are multiple modes, try to understand why. If you can identify a reason for separate modes (for example, women and men typically have heart attacks at different ages), it may be a good idea to split the data into separate groups.
  - If there are any clear outliers, you should point them out. If you are reporting the mean and standard deviation, report them with the outliers present and with the outliers removed. The differences may be revealing. (Of course, the median and IQR won't be affected very much by the outliers.)

STEP-BY-STEP EXAMPLE — Summarizing a distribution

One of the authors owned a 1989 Nissan Maxima for 8 years. Being a statistician, he recorded the car's fuel efficiency (in mpg) each time he filled the tank. He wanted to know what fuel efficiency to expect as "ordinary" for his car. (Hey, he's a statistician. What would you expect?[9]) Knowing this, he was able to predict when he'd need to fill the tank again and to notice if the fuel efficiency suddenly got worse, which could be a sign of trouble.

Question: How would you describe the distribution of *Fuel efficiency* for this car?

Plan State what you want to find out.

I want to summarize the distribution of Nissan Maxima fuel efficiency.

Variable Identify the variable and report the W's.

The data are the fuel efficiency values in miles per gallon for the first 100 fill-ups of a 1989 Nissan Maxima between 1989 and 1992.

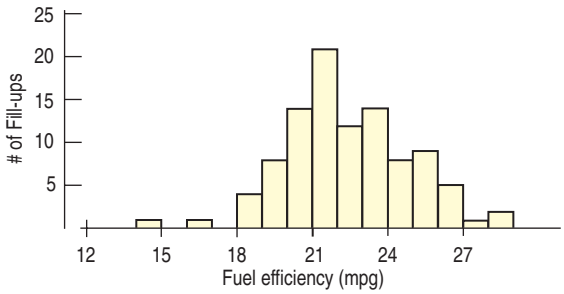
Be sure to check the appropriate condition.

✓ **Quantitative Data Condition:** The fuel efficiencies are quantitative with units of miles per gallon. Histograms and boxplots are appropriate displays for displaying the distribution. Numerical summaries are appropriate as well.

[9] He also recorded the time of day, temperature, price of gas, and phase of the moon. (OK, maybe not phase of the moon.) His data are on the DVD.

SHOW

Mechanics Make a histogram and boxplot. Based on the shape, choose appropriate numerical summaries.



A histogram of the data shows a fairly symmetric distribution with a low outlier.

REALITY CHECK A value of 22 mpg seems reasonable for such a car. The spread is reasonable, although the range looks a bit large.

Count	100
Mean	22.4 mpg
StdDev	2.45
Q1	20.8
Median	22.0
Q3	24.0
IQR	3.2

The mean and median are close, so the outlier doesn't seem to be a problem. I can use the mean and standard deviation.

TELL

Conclusion Summarize and interpret your findings in context. Be sure to discuss the distribution's shape, center, spread, and unusual features (if any).

The distribution of mileage is unimodal and roughly symmetric with a mean of 22.4 mpg. There is a low outlier that should be investigated, but it does not influence the mean very much. The standard deviation suggests that from tankful to tankful, I can expect the car's fuel economy to differ from the mean by an average of about 2.45 mpg.

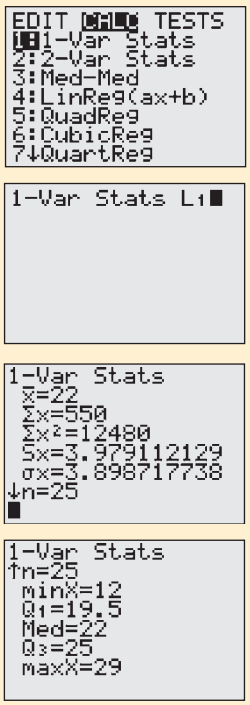
Are my statistics "right"? When you calculate a mean, the computation is clear: You sum all the values and divide by the sample size. You may round your answer less or more than someone else (we recommend one more decimal place than the data), but all books and technologies agree on how to find the mean. Some statistics, however, are more problematic. For example we've already pointed out that methods of finding quartiles differ.

Differences in numeric results can also arise from decisions in the middle of calculations. For example, if you round off your value for the mean before you calculate the sum of squared deviations, your standard deviation probably won't agree with a computer program that calculates using many decimal places. (We do recommend that you do calculations using as many digits as you can to minimize this effect.)

Don't be overly concerned with these discrepancies, especially if the differences are small. They don't mean that your answer is "wrong," and they usually won't change any conclusion you might draw about the data. Sometimes (in footnotes and in the answers in the back of the book) we'll note alternative results, but we could never list all the possible values, so we'll rely on your common sense to focus on the meaning rather than on the digits. Remember: Answers are sentences!

TI Tips

Calculating the statistics



Your calculator can easily find all the numerical summaries of data. To try it out, you simply need a set of values in one of your datalists. We'll illustrate using the boys' agility test results from this chapter's earlier TI Tips (still in `L1`), but you can use any data currently stored in your calculator.

- Under the `STAT CALC` menu, select `1-Var Stats` and hit `ENTER`.
- Specify the location of your data, creating a command like `1-Var Stats L1`.
- Hit `ENTER` again.

Voilà! Everything you wanted to know, and more. Among all of the information shown, you are primarily interested in these statistics: $\bar{x}$ (the mean), `Sx` (the standard deviation), `n` (the count), and—scrolling down—`minX` (the smallest datum), Q_1 (the first quartile), `Med` (the median), Q_3 (the third quartile), and `maxX` (the largest datum).

Sorry, but the TI doesn't explicitly tell you the range or the IQR. Just subtract: IQR = $Q_3 - Q_1 = 25 - 19.5 = 5.5$. What's the range?

By the way, if the data come as a frequency table with the values stored in, say, `L4` and the corresponding frequencies in `L5`, all you have to do is ask for `1-Var Stats L4,L5`.

WHAT CAN GO WRONG?

A data display should tell a story about the data. To do that, it must speak in a clear language, making plain what variable is displayed, what any axis shows, and what the values of the data are. And it must be consistent in those decisions.

A display of quantitative data can go wrong in many ways. The most common failures arise from only a few basic errors:

- **Don't make a histogram of a categorical variable.** Just because the variable contains numbers doesn't mean that it's quantitative. Here's a histogram of the insurance policy numbers of some workers. It's not very informative because the policy numbers are just labels. A histogram or stem-and-leaf display of a categorical variable makes no sense. A bar chart or pie chart would be more appropriate.
- **Don't look for shape, center, and spread of a bar chart.** A bar chart showing the sizes of the piles displays the distribution of a categorical variable, but the bars could be arranged in any order left to right. Concepts like symmetry, center, and spread make sense only for quantitative variables.

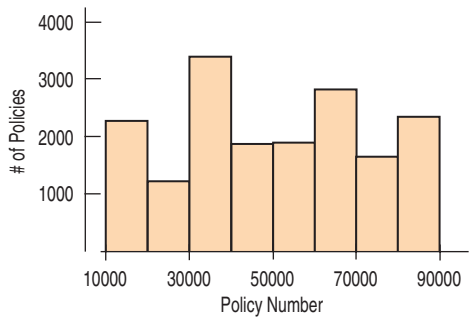
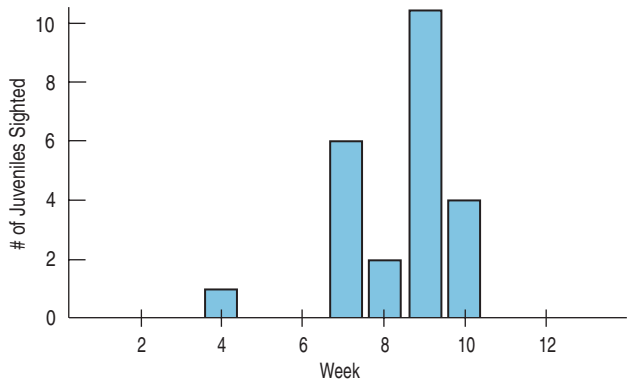


FIGURE 4.14
It's not appropriate to display these data with a histogram.

(*continued*)

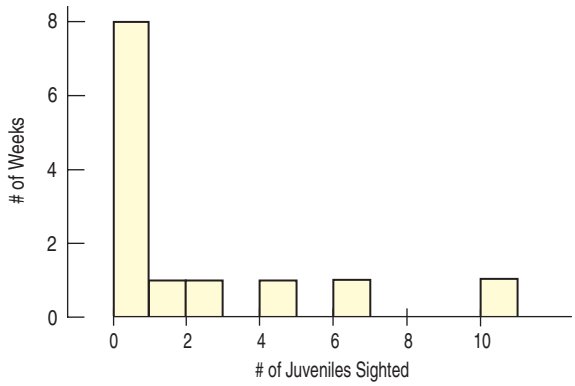
- **Don't use bars in every display—save them for histograms and bar charts.** In a bar chart, the bars indicate how many cases of a categorical variable are piled in each category. Bars in a histogram indicate the number of cases piled in each interval of a quantitative variable. In both bar charts and histograms, the bars represent counts of data values. Some people create other displays that use bars to represent individual data values. Beware: Such graphs are neither bar charts nor histograms. For example, a student was asked to make a histogram from data showing the number of juvenile bald eagles seen during each of the 13 weeks in the winter of 2003–2004 at a site in Rock Island, IL. Instead, he made this plot:

FIGURE 4.15
This isn't a histogram or a bar chart. It's an ill-conceived graph that uses bars to represent individual data values (number of eagles sighted) week by week.



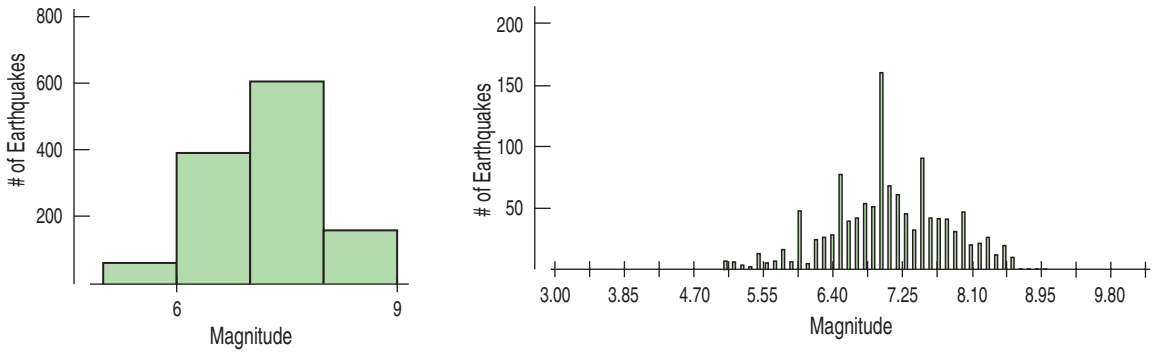
Look carefully. That's not a histogram. A histogram shows *What* we've measured along the horizontal axis and counts of the associated *Who*'s represented as bar heights. This student has it backwards: He used bars to show counts of birds for each week.[10] We need counts of weeks. A correct histogram should have a tall bar at "0" to show there were many weeks when no eagles were seen, like this:

FIGURE 4.16
A histogram of the eagle-sighting data shows the number of weeks in which different counts of eagles occurred. This display shows the distribution of juvenile-eagle sightings.



- **Choose a bin width appropriate to the data.** Computer programs usually do a pretty good job of choosing histogram bin widths. Often there's an easy way to adjust the width, sometimes interactively. Here are the tsunami earthquakes with two (rather extreme) choices for the bin size:

[10] Edward Tufte, in his book *The Visual Display of Quantitative Information*, proposes that graphs should have a high data-to-ink ratio. That is, we shouldn't waste a lot of ink to display a single number when a dot would do the job.



The task of summarizing a quantitative variable is relatively simple, and there is a simple path to follow. However, you need to watch out for certain features of the data that make summarizing them with a number dangerous. Here's some advice:

- **Don't forget to do a reality check.** Don't let the computer or calculator do your thinking for you. Make sure the calculated summaries make sense. For example, does the mean look like it is in the center of the histogram? Think about the spread: An IQR of 50 mpg would clearly be wrong for gas mileage. And no measure of spread can be negative. The standard deviation can take the value 0, but only in the very unusual case that all the data values equal the same number. If you see an IQR or standard deviation equal to 0, it's probably a sign that something's wrong with the data.
- **Don't forget to sort the values before finding the median or percentiles.** It seems obvious, but when you work by hand, it's easy to forget to sort the data first before counting in to find medians, quartiles, or other percentiles. Don't report that the median of the five values 194, 5, 1, 17, and 893 is 1 just because 1 is the middle number.
- **Don't worry about small differences when using different methods.** Finding the 10th percentile or the lower quartile in a data set sounds easy enough. But it turns out that the definitions are not exactly clear. If you compare different statistics packages or calculators, you may find that they give slightly different answers for the same data. These differences, though, are unlikely to be important in interpreting the data, the quartiles, or the IQR, so don't let them worry you.
- **Don't compute numerical summaries of a categorical variable.** Neither the mean zip code nor the standard deviation of social security numbers is meaningful. If the variable is categorical, you should instead report summaries such as percentages of individuals in each category. It is easy to make this mistake when using technology to do the summaries for you. After all, the computer doesn't care what the numbers mean.

Gold Card Customers—Regions National Banks

Month	April 2007	May 2007
Average Zip Code	45,034.34	38,743.34

- **Don't report too many decimal places.** Statistical programs and calculators often report a ridiculous number of digits. A general rule for numerical summaries is to report one or two more digits than the number of digits in the data. For example, earlier we saw a dotplot of the Kentucky Derby race times. The mean and standard deviation of those times could be reported as:

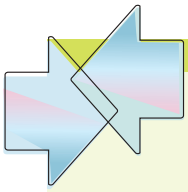
$$\bar{y} = 130.63401639344262\,\text{sec} \qquad s = 13.66448201942662\,\text{sec}$$

But we knew the race times only to the nearest quarter second, so the extra digits are meaningless.

- **Don't round in the middle of a calculation.** Don't *report* too many decimal places, but it's best not to do any rounding until the end of your calculations. Even though you might report the mean of the earthquakes as 7.08, it's really 7.08339. Use the more precise number in your calculations if you're finding the standard deviation by hand—or be prepared to see small differences in your final result.

(*continued*)

- **Watch out for multiple modes.** The summaries of the Kentucky Derby times are meaningless for another reason. As we saw in the dotplot, the Derby was initially a longer race. It would make much more sense to report that the old 1.5 mile Derby had a mean time of 159.6 seconds, while the current Derby has a mean time of 124.6 seconds. If the distribution has multiple modes, consider separating the data into different groups and summarizing each group separately.
- **Beware of outliers.** The median and IQR are resistant to outliers, but the mean and standard deviation are not. To help spot outliers . . .
- **Don't forget to: Make a picture (make a picture, make a picture).** The sensitivity of the mean and standard deviation to outliers is one reason you should always make a picture of the data. Summarizing a variable with its mean and standard deviation when you have not looked at a histogram or dotplot to check for outliers or skewness invites disaster. You may find yourself drawing absurd or dangerously wrong conclusions about the data. And, of course, you should demand no less of others. Don't accept a mean and standard deviation blindly without some evidence that the variable they summarize is unimodal, symmetric, and free of outliers.



CONNECTIONS

Distributions of quantitative variables, like those of categorical variables, show the possible values and their relative frequencies. A histogram shows the distribution of values in a quantitative variable with adjacent bars. Don't confuse histograms with bar charts, which display categorical variables. For categorical data, the mode is the category with the biggest count. For quantitative data, modes are peaks in the histogram.

The shape of the distribution of a quantitative variable is an important concept in most of the subsequent chapters. We will be especially interested in distributions that are unimodal and symmetric.

In addition to their shape, we summarize distributions with center and spread, usually pairing a measure of center with a measure of spread: median with IQR and mean with standard deviation. We favor the mean and standard deviation when the shape is unimodal and symmetric, but choose the median and IQR for skewed distributions or when there are outliers we can't otherwise set aside.

WHAT HAVE WE LEARNED?

We've learned how to make a picture of quantitative data to help us see the story the data have to *Tell*.

- We can display the distribution of quantitative data with a *histogram*, a *stem-and-leaf* display, or a *dotplot*.
- We *Tell* what we see about the distribution by talking about *shape*, *center*, *spread*, and any *unusual features*.

We've learned how to summarize distributions of quantitative variables numerically.

- Measures of center for a distribution include the median and the mean.

  We write the formula for the mean as $\bar{y} = \frac{\sum y}{n}$.
- Measures of spread include the range, IQR, and standard deviation.

  The standard deviation is computed as $s = \sqrt{\frac{\sum (y - \bar{y})^2}{n - 1}}$.

  The median and IQR are not usually given as formulas.

- We'll report the median and IQR when the distribution is skewed. If it's symmetric, we'll summarize the distribution with the mean and standard deviation (and possibly the median and IQR as well). Always pair the median with the IQR and the mean with the standard deviation.

We've learned to *Think* about the type of variable we're summarizing.

- All the methods of this chapter assume that the data are quantitative.
- The **Quantitative Data Condition** serves as a check that the data are, in fact, quantitative. One good way to be sure is to know the measurement units. You'll want those as part of the *Think* step of your answers.

Terms

Distribution 44. The distribution of a quantitative variable slices up all the possible values of the variable into equal-width bins and gives the number of values (or counts) falling into each bin.

Histogram (relative frequency histogram) 45. A histogram uses adjacent bars to show the distribution of a quantitative variable. Each bar represents the frequency (or relative frequency) of values falling in each bin.

Gap 45. A region of the distribution where there are no values.

Stem-and-leaf display 47. A stem-and-leaf display shows quantitative data values in a way that sketches the distribution of the data. It's best described in detail by example.

Dotplot 49. A dotplot graphs a dot for each case against a single axis.

Shape 49. To describe the shape of a distribution, look for
- single vs. multiple modes.
- symmetry vs. skewness.
- outliers and gaps.

Center 52, 58. The place in the distribution of a variable that you'd point to if you wanted to attempt the impossible by summarizing the entire distribution with a single number. Measures of center include the mean and median.

Spread 54, 61. A numerical summary of how tightly the values are clustered around the center. Measures of spread include the IQR and standard deviation.

Mode 49. A hump or local high point in the shape of the distribution of a variable. The apparent location of modes can change as the scale of a histogram is changed.

Unimodal (Bimodal) 50. Having one mode. This is a useful term for describing the shape of a histogram when it's generally mound-shaped. Distributions with two modes are called **bimodal.** Those with more than two are **multimodal.**

Uniform 50. A distribution that's roughly flat is said to be uniform.

Symmetric 50. A distribution is symmetric if the two halves on either side of the center look approximately like mirror images of each other.

Tails 50. The tails of a distribution are the parts that typically trail off on either side. Distributions can be characterized as having long tails (if they straggle off for some distance) or short tails (if they don't).

Skewed 50. A distribution is skewed if it's not symmetric and one tail stretches out farther than the other. Distributions are said to be **skewed left** when the longer tail stretches to the left, and **skewed right** when it goes to the right.

Outliers 51. Outliers are extreme values that don't appear to belong with the rest of the data. They may be unusual values that deserve further investigation, or they may be just mistakes; there's no obvious way to tell. Don't delete outliers automatically—you have to think about them. Outliers can affect many statistical analyses, so you should always be alert for them.

Median 52. The median is the middle value, with half of the data above and half below it. If n is even, it is the average of the two middle values. It is usually paired with the IQR.

Range 54. The difference between the lowest and highest values in a data set. $Range = max - min$.

Quartile 54. The lower quartile (Q1) is the value with a quarter of the data below it. The upper quartile (Q3) has three quarters of the data below it. The median and quartiles divide data into four parts with equal numbers of data values.

Interquartile range (IQR) 55. The IQR is the difference between the first and third quartiles. $IQR = Q3 - Q1$. It is usually reported along with the median.

Percentile 55. The ith percentile is the number that falls above i% of the data.

5-Number Summary 56. The 5-number summary of a distribution reports the minimum value, Q1, the median, Q3, and the maximum value.

Mean 58. The mean is found by summing all the data values and dividing by the count:

$$\bar{y} = \frac{Total}{n} = \frac{\sum y}{n}.$$

It is usually paired with the standard deviation.

Resistant 59. A calculated summary is said to be resistant if outliers have only a small effect on it.

Variance 61. The variance is the sum of squared deviations from the mean, divided by the count minus 1:

$$s^2 = \frac{\sum (y - \bar{y})^2}{n - 1}.$$

It is useful in calculations later in the book.

Standard deviation 61. The standard deviation is the square root of the variance:

$$s = \sqrt{\frac{\sum (y - \bar{y})^2}{n - 1}}$$

It is usually reported along with the mean.

Skills

THINK

- Be able to identify an appropriate display for any quantitative variable.
- Be able to guess the shape of the distribution of a variable by knowing something about the data.
- Be able to select a suitable measure of center and a suitable measure of spread for a variable based on information about its distribution.
- Know the basic properties of the median: The median divides the data into the half of the data values that are below the median and the half that are above.
- Know the basic properties of the mean: The mean is the point at which the histogram balances.
- Know that the standard deviation summarizes how spread out all the data are around the mean.
- Understand that the median and IQR resist the effects of outliers, while the mean and standard deviation do not.
- Understand that in a skewed distribution, the mean is pulled in the direction of the skewness (toward the longer tail) relative to the median.

SHOW

- Know how to display the distribution of a quantitative variable with a stem-and-leaf display (drawn by hand for smaller data sets), a dotplot, or a histogram (made by computer for larger data sets).
- Know how to compute the mean and median of a set of data.
- Know how to compute the standard deviation and IQR of a set of data.

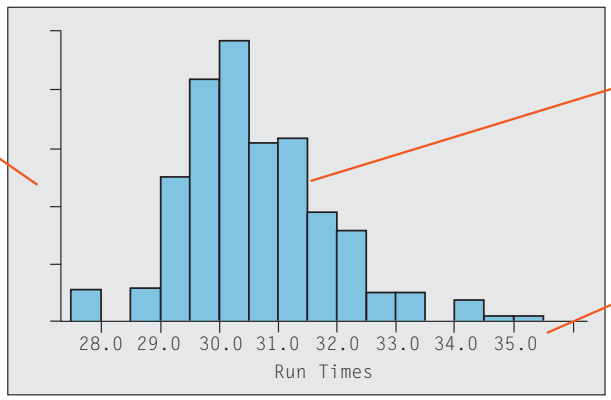
TELL

- Be able to describe the distribution of a quantitative variable in terms of its shape, center, and spread.
- Be able to describe any anomalies or extraordinary features revealed by the display of a variable.
- Know how to describe summary measures in a sentence. In particular, know that the common measures of center and spread have the same units as the variable that they summarize, and should be described in those units.
- Be able to describe the distribution of a quantitative variable with a description of the shape of the distribution, a numerical measure of center, and a numerical measure of spread. Be sure to note any unusual features, such as outliers, too.

DISPLAYING AND SUMMARIZING QUANTITATIVE VARIABLES ON THE COMPUTER

Almost any program that displays data can make a histogram, but some will do a better job of determining where the bars should start and how they should partition the span of the data.

The vertical scale may be counts or proportions. Sometimes it isn't clear which. But the shape of the histogram is the same either way.



Most packages choose the number of bars for you automatically. Often you can adjust that choice.

The axis should be clearly labeled so you can tell what "pile" each bar represents. You should be able to tell the lower and upper bounds of each bar.

Many statistics packages offer a prepackaged collection of summary measures. The result might look like this:

Variable: W eight
N = 234
Mean = 143.3 Median = 139
St. Dev = 11.1 IQR = 14

Alternatively, a package might make a table for several variables and summary measures:

Variable	N	mean	median	stdev	IQR
Weight	234	143.3	139	11.1	14
Height	234	68.3	68.1	4.3	5
Score	234	86	88	9	5

A S ***Case Study:*** **Describing Distribution Shapes.** Who's safer in a crash—passengers or the driver? Investigate with your statistics package.

It is usually easy to read the results and identify each computed summary. You should be able to read the summary statistics produced by any computer package.

Packages often provide many more summary statistics than you need. Of course, some of these may not be appropriate when the data are skewed or have outliers. It is your responsibility to check a histogram or stem-and-leaf display and decide which summary statistics to use.

It is common for packages to report summary statistics to many decimal places of "accuracy." Of course, it is rare data that have such accuracy in the original measurements. The ability to calculate to six or seven digits beyond the decimal point doesn't mean that those digits have any meaning. Generally it's a good idea to round these values, allowing perhaps one more digit of precision than was given in the original data.

Displays and summaries of quantitative variables are among the simplest things you can do in most statistics packages.

As practice, work on problems # 1, 3, 6, 10, 14, 15, 16, 26, 28, 29, and 30.

PART I

REVIEW OF PART I

Exploring and Understanding Data

Quick Review

It's time to put it all together. Real data don't come tagged with instructions for use. So let's step back and look at how the key concepts and skills we've seen work together. This brief list and the review exercises that follow should help you check your understanding of Statistics so far.

- We treat data two ways: as categorical and as quantitative.
- To describe categorical data:
  - Make a picture. Bar graphs work well for comparing counts in categories.
  - Summarize the distribution with a table of counts or relative frequencies (percents) in each category.
  - Pie charts and segmented bar charts display divisions of a whole.
  - Compare distributions with plots side by side.
  - Look for associations between variables by comparing marginal and conditional distributions.
- To describe quantitative data:
  - Make a picture. Use histograms, boxplots, stem-and-leaf displays, or dotplots. Stem-and-leafs are great when working by hand and good for small data sets. Histograms are a good way to see the distribution. Boxplots are best for comparing several distributions.
  - Describe distributions in terms of their shape, center, and spread, and note any unusual features such as gaps or outliers.
  - The shape of most distributions you'll see will likely be uniform, unimodal, or bimodal. It may be multimodal. If it is unimodal, then it may be symmetric or skewed.
  - A 5-number summary makes a good numerical description of a distribution: min, Q1, median, Q3, and max.
  - If the distribution is skewed, be sure to include the median and interquartile range (IQR) when you describe its center and spread.
  - A distribution that is severely skewed may benefit from re-expressing the data. If it is skewed to the high end, taking logs often works well.
  - If the distribution is unimodal and symmetric, describe its center and spread with the mean and standard deviation.
  - Use the standard deviation as a ruler to tell how unusual an observed value may be, or to compare or combine measurements made on different scales.
  - Shifting a distribution by adding or subtracting a constant affects measures of position but not measures of spread. Rescaling by multiplying or dividing by a constant affects both.
  - When a distribution is roughly unimodal and symmetric, a Normal model may be useful. For Normal models, the 68–95–99.7 Rule is a good rule of thumb.
  - If the Normal model fits well (check a histogram or Normal probability plot), then Normal percentile tables or functions found in most statistics technology can provide more detailed values.

Need more help with some of this? It never hurts to reread sections of the chapters! And in the following pages we offer you more opportunities[1] to review these concepts and skills.

The exercises that follow use the concepts and skills you've learned in the first six chapters. To be more realistic and more useful for your review, they don't tell you which of the concepts or methods you need. But neither will the exam.

[1] If you doubted that we are teachers, this should convince you. Only a teacher would call additional homework exercises "opportunities."

REVIEW EXERCISES

1. **Bananas.** Here are the prices (in cents per pound) of bananas reported from 15 markets surveyed by the U.S. Department of Agriculture.

51	52	45
48	53	52
50	49	52
48	43	46
45	42	50

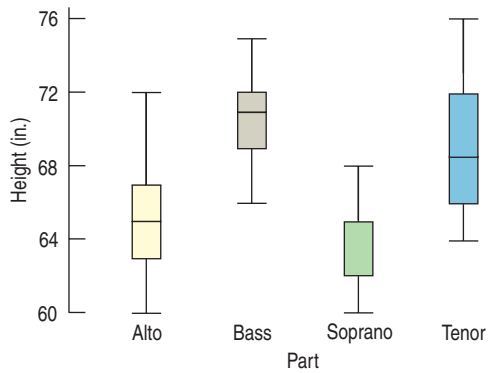
a) Display these data with an appropriate graph.
b) Report appropriate summary statistics.
c) Write a few sentences about this distribution.

2. **Prenatal care.** Results of a 1996 American Medical Association report about the infant mortality rate for twins carried for the full term of a normal pregnancy are shown on the next page, broken down by the level of prenatal care the mother had received.

Full-Term Pregnancies, Level of Prenatal Care	Infant Mortality Rate Among Twins (deaths per thousand live births)
Intensive	5.4
Adequate	3.9
Inadequate	6.1
Overall	5.1

a) Is the overall rate the average of the other three rates? Should it be? Explain.
b) Do these results indicate that adequate prenatal care is important for pregnant women? Explain.
c) Do these results suggest that a woman pregnant with twins should be wary of seeking too much medical care? Explain.

3. **Singers.** The boxplots shown display the heights (in inches) of 130 members of a choir.

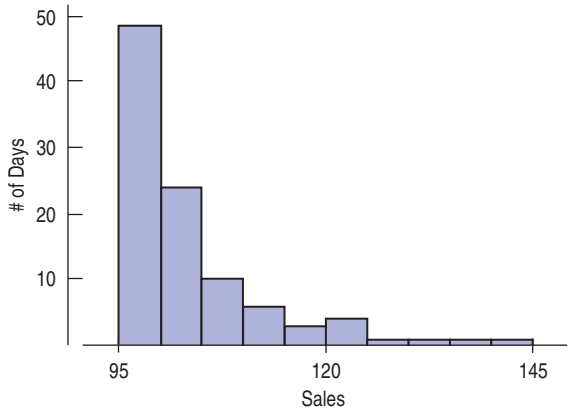


a) It appears that the median height for sopranos is missing, but actually the median and the upper quartile are equal. How could that happen?
b) Write a few sentences describing what you see.

4. **Dialysis.** In a study of dialysis, researchers found that "of the three patients who were currently on dialysis, 67% had developed blindness and 33% had their toes amputated." What kind of display might be appropriate for these data? Explain.

5. **Beanstalks.** Beanstalk Clubs are social clubs for very tall people. To join, a man must be over 6′2″ tall, and a woman over 5′10″. The National Health Survey suggests that heights of adults may be Normally distributed, with mean heights of 69.1″ for men and 64.0″ for women. The respective standard deviations are 2.8″ and 2.5″.
a) You are probably not surprised to learn that men are generally taller than women, but what does the greater standard deviation for men's heights indicate?
b) Who are more likely to qualify for Beanstalk membership, men or women? Explain.

6. **Bread.** Clarksburg Bakery is trying to predict how many loaves to bake. In the last 100 days, they have sold between 95 and 140 loaves per day. Here is a histogram of the number of loaves they sold for the last 100 days.



a) Describe the distribution.
b) Which should be larger, the mean number of sales or the median? Explain.
c) Here are the summary statistics for Clarksburg Bakery's bread sales. Use these statistics and the histogram above to create a boxplot. You may approximate the values of any outliers.

Summary of Sales	
Median	100
Min	95
Max	140
25th %tile	97
75th %tile	105.5

d) For these data, the mean was 103 loaves sold per day, with a standard deviation of 9 loaves. Do these statistics suggest that Clarksburg Bakery should expect to sell between 94 and 112 loaves on about 68% of the days? Explain.

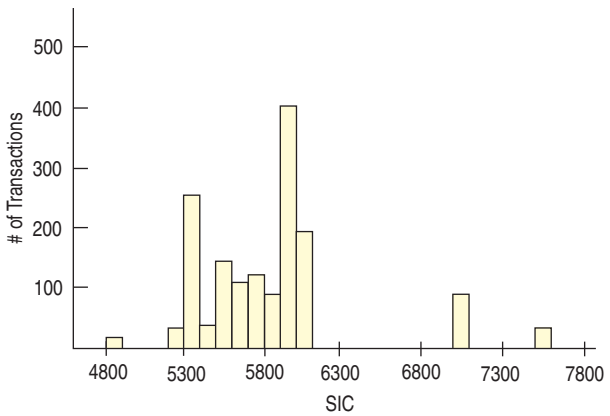
7. **State University.** Public relations staff at State U. collected data on people's opinions of various colleges and universities in their state. They phoned 850 local residents. After identifying themselves, the callers asked the survey participants their ages, whether they had attended college, and whether they had a favorable opinion of the university. The official report to the university's directors claimed that, in general, people had very favorable opinions about their university.
a) Identify the W's of these data.
b) Identify the variables, classify each as categorical or quantitative, and specify units if relevant.
c) Are you confident about the report's conclusion? Explain.

8. **Acid rain.** Based on long-term investigation, researchers have suggested that the acidity (pH) of rainfall

in the Shenandoah Mountains can be described by the Normal model $N(4.9, 0.6)$.
a) Draw and carefully label the model.
b) What percent of storms produce rainfall with pH over 6?
c) What percent of storms produce rainfall with pH under 4?
d) The lower the pH, the more acidic the rain. What is the pH level for the most acidic 20% of all storms?
e) What is the pH level for the least acidic 5% of all storms?
f) What is the IQR for the pH of rainfall?

9. Fraud detection. A credit card bank is investigating the incidence of fraudulent card use. The bank suspects that the type of product bought may provide clues to the fraud. To examine this situation, the bank looks at the Standard Industrial Code (SIC) of the business related to the transaction. This is a code that was used by the U.S. Census Bureau and Statistics Canada to identify the type of every registered business in North America.[2] For example, 1011 designates Meat and Meat Products (except Poultry), 1012 is Poultry Products, 1021 is Fish Products, 1031 is Canned and Preserved Fruits and Vegetables, and 1032 is Frozen Fruits and Vegetables.

A company intern produces the following histogram of the SIC codes for 1536 transactions:



He also reports that the mean SIC is 5823.13 with a standard deviation of 488.17.
a) Comment on any problems you see with the use of the mean and standard deviation as summary statistics.
b) How well do you think the Normal model will work on these data? Explain.

10. Streams. As part of the course work, a class at an upstate NY college collects data on streams each year. Students record a number of biological, chemical, and physical variables, including the stream name, the substrate of the stream (*limestone, shale, or mixed*), the pH, the temperature (°C), and the BCI, a measure of biological diversity.

Group	Count	%
Limestone	77	44.8
Mixed	26	15.1
Shale	69	40.1

a) Name each variable, indicating whether it is categorical or quantitative, and giving the units if available.
b) These streams have been classified according to their substrate—the composition of soil and rock over which they flow—as summarized in the table. What kind of graph might be used to display these data?

T **11. Cramming.** One Thursday, researchers gave students enrolled in a section of basic Spanish a set of 50 new vocabulary words to memorize. On Friday the students took a vocabulary test. When they returned to class the following Monday, they were retested—without advance warning. Both sets of test scores for the 28 students are shown below.

Fri	Mon	Fri	Mon
42	36	50	47
44	44	34	34
45	46	38	31
48	38	43	40
44	40	39	41
43	38	46	32
41	37	37	36
35	31	40	31
43	32	41	32
48	37	48	39
43	41	37	31
45	32	36	41
47	44		

a) Create a graphical display to compare the two distributions of scores.
b) Write a few sentences about the scores reported on Friday and Monday.
c) Create a graphical display showing the distribution of the *changes* in student scores.
d) Describe the distribution of changes.

12. Computers and Internet. A U.S. Census Bureau report (August 2000, *Current Population Survey*) found that 51.0% of homes had a personal computer and 41.5% had access to the Internet. A newspaper concluded that 92.5% of homes had either a computer or access to the Internet. Do you agree? Explain.

13. Let's play cards. You pick a card from a deck (see description in Chapter 11) and record its denomination (7, say) and its suit (maybe spades).
a) Is the variable *suit* categorical or quantitative?
b) Name a game you might be playing for which you would consider the variable *denomination* to be categorical. Explain.
c) Name a game you might be playing for which you would consider the variable *denomination* to be quantitative. Explain.

T **14. Accidents.** In 2001, Progressive Insurance asked customers who had been involved in auto accidents how far they were from home when the accident happened. The data are summarized in the table.

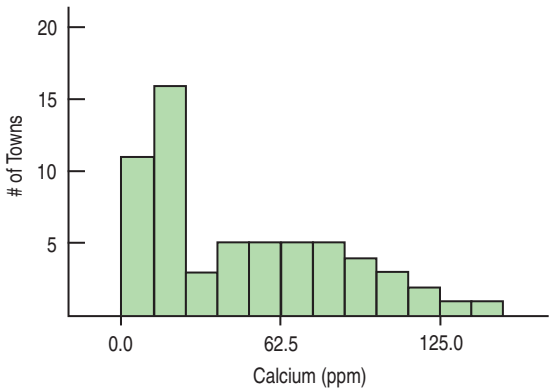
[2] Since 1997 the SIC has been replaced by the NAICS, a code of six letters.

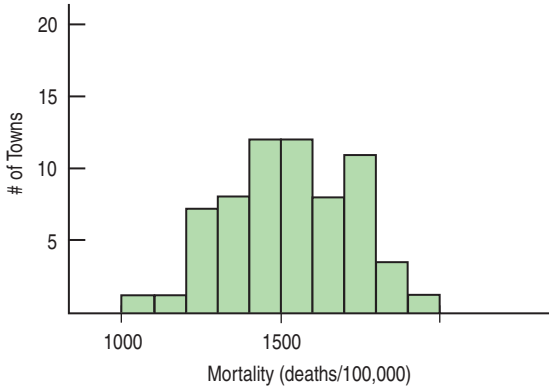
Miles from Home	% of Accidents
Less than 1	23
1 to 5	29
6 to 10	17
11 to 15	8
16 to 20	6
Over 20	17

a) Create an appropriate graph of these data.
b) Do these data indicate that driving near home is particularly dangerous? Explain.

15. Hard water. In an investigation of environmental causes of disease, data were collected on the annual mortality rate (deaths per 100,000) for males in 61 large towns in England and Wales. In addition, the water hardness was recorded as the calcium concentration (parts per million, ppm) in the drinking water.

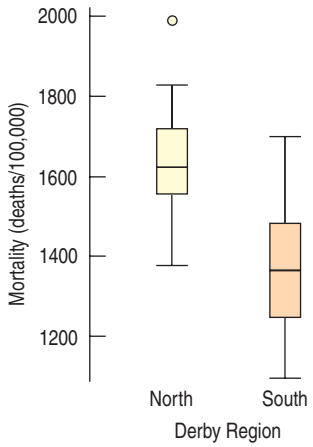
a) What are the variables in this study? For each, indicate whether it is quantitative or categorical and what the units are.
b) Here are histograms of calcium concentration and mortality. Describe the distributions of the two variables.





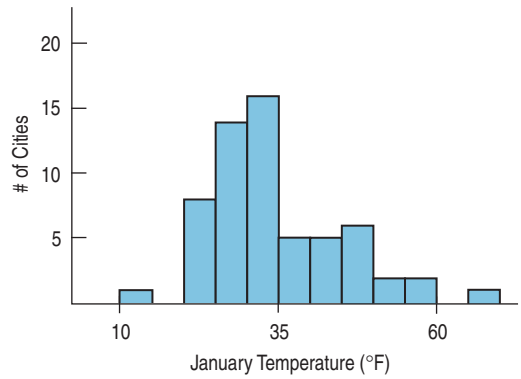
16. Hard water II. The data set from England and Wales also notes for each town whether it was south or north of Derby. Here are some summary statistics and a comparative boxplot for the two regions.

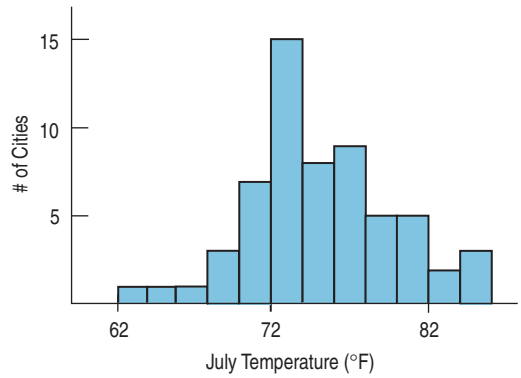
Summary of Mortality				
Group	Count	Mean	Median	StdDev
North	34	1631.59	1631	138.470
South	27	1388.85	1369	151.114



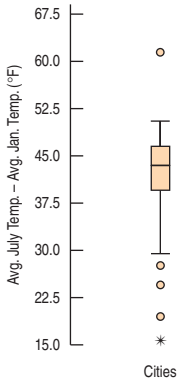
a) What is the overall mean mortality rate for the two regions?
b) Do you see evidence of a difference in mortality rates? Explain.

17. Seasons. Average daily temperatures in January and July for 60 large U.S. cities are graphed in the histograms below.



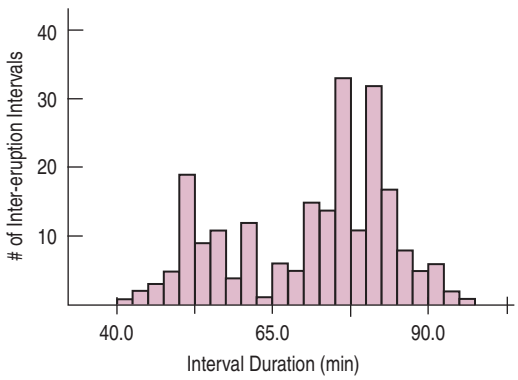


a) What aspect of these histograms makes it difficult to compare the distributions?
b) What differences do you see between the distributions of January and July average temperatures?



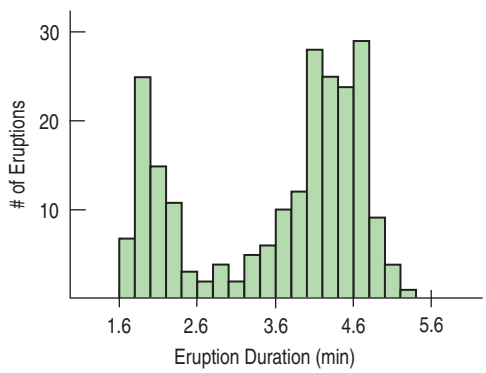
c) Differences in temperatures (July–January) for each of the cities are displayed in the boxplot above. Write a few sentences describing what you see.

18. Old Faithful. It is a common belief that Yellowstone's most famous geyser erupts once an hour at very predictable intervals. The histogram below shows the time gaps (in minutes) between 222 successive eruptions. Describe this distribution.

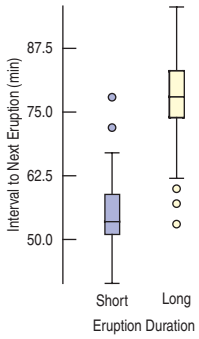


19. Old Faithful? Does the duration of an eruption have an effect on the length of time that elapses before the next eruption?

a) The histogram below shows the duration (in minutes) of those 222 eruptions. Describe this distribution.



b) Explain why it is not appropriate to find summary statistics for this distribution.
c) Let's classify the eruptions as "long" or "short," depending upon whether or not they last at least 3 minutes. Describe what you see in the comparative boxplots.



20. Teen drivers. In its *Traffic Safety Facts 2005*, the National Highway Traffic Safety Administration reported that 6.3% of licensed drivers were between the ages of 15 and 20, yet this age group was behind the wheel in 15.9% of all fatal crashes. Use these statistics to explain the concept of independence.

T **21. Liberty's nose.** Is the Statue of Liberty's nose too long? Her nose measures, 4′6″, but she is a large statue, after all. Her arm is 42 feet long. That means her arm is 42/45 = 9.3 times as long as her nose. Is that a reasonable ratio? Shown in the table are arm and nose lengths of 18 girls in a Statistics class, and the ratio of arm-to-nose length for each.

Arm (cm)	Nose (cm)	Arm/Nose Ratio
73.8	5.0	14.8
74.0	4.5	16.4
69.5	4.5	15.4
62.5	4.7	13.3
68.6	4.4	15.6
64.5	4.8	13.4
68.2	4.8	14.2
63.5	4.4	14.4
63.5	5.4	11.8
67.0	4.6	14.6
67.4	4.4	15.3
70.7	4.3	16.4
69.4	4.1	16.9
71.7	4.5	15.9
69.0	4.4	15.7
69.8	4.5	15.5
71.0	4.8	14.8
71.3	4.7	15.2

a) Make an appropriate plot and describe the distribution of the ratios.
b) Summarize the ratios numerically, choosing appropriate measures of center and spread.
c) Is the ratio of 9.3 for the Statue of Liberty unrealistically low? Explain.

T **22. Winter Olympics 2006 speed skating.** The top 25 women's 500-m speed skating times are listed in the table below:

Skater	Country	Time
Svetlana Zhurova	Russia	76.57
Wang Manli	China	76.78
Hui Ren	China	76.87
Tomomi Okazaki	Japan	76.92
Lee Sang-Hwa	South Korea	77.04
Jenny Wolf	Germany	77.25
Wang Beixing	China	77.27
Sayuri Osuga	Japan	77.39
Sayuri Yoshii	Japan	77.43
Chiara Simionato	Italy	77.68
Jennifer Rodriguez	United States	77.70
Annette Gerritsen	Netherlands	78.09
Xing Aihua	China	78.35
Sanne van der Star	Netherlands	78.59
Yukari Watanabe	Japan	78.65
Shannon Rempel	Canada	78.85
Amy Sannes	United States	78.89
Choi Seung-Yong	South Korea	79.02
Judith Hesse	Germany	79.03
Kim You-Lim	South Korea	79.25
Kerry Simpson	Canada	79.34
Krisy Myers	Canada	79.43
Elli Ochowicz	United States	79.48
Pamela Zoellner	Germany	79.56
Lee Bo-Ra	South Korea	79.73

a) The mean finishing time was 78.21 seconds, with a standard deviation of 1.03 second. If the Normal model is appropriate, what percent of the times should be within 0.5 second of 78.21?
b) What percent of the times actually fall within this interval?
c) Explain the discrepancy between a and b.

23. Sample. A study in South Africa focusing on the impact of health insurance identified 1590 children at birth and then sought to conduct follow-up health studies 5 years later. Only 416 of the original group participated in the 5-year follow-up study. This made researchers concerned that the follow-up group might not accurately resemble the total group in terms of health insurance. The table in the next column summarizes the two groups by race and by presence of medical insurance when the child was born. Carefully explain how this study demonstrates Simpson's paradox. (*Birth to Ten Study*, Medical Research Council, South Africa)

		Number (%) Insured	
		Follow-up	**Not traced**
Race	**Black**	36 of 404 (8.9%)	91 of 1048 (8.7%)
	White	10 of 12 (83.3%)	104 of 126 (82.5%)
	Overall	46 of 416 (11.1%)	195 of 1174 (16.6%)

24. Sluggers. Roger Maris's 1961 home run record stood until Mark McGwire hit 70 in 1998. Listed below are the home run totals for each season McGwire played. Also listed are Babe Ruth's home run totals.

McGwire: 3*, 49, 32, 33, 39, 22, 42, 9*, 9*, 39, 52, 58, 70, 65, 32*, 29*

Ruth: 54, 59, 35, 41, 46, 25, 47, 60, 54, 46, 49, 46, 41, 34, 22

a) Find the 5-number summary for McGwire's career.
b) Do any of his seasons appear to be outliers? Explain.
c) McGwire played in only 18 games at the end of his first big league season, and missed major portions of some other seasons because of injuries to his back and knees. Those seasons might not be representative of his abilities. They are marked with asterisks in the list above. Omit these values and make parallel boxplots comparing McGwire's career to Babe Ruth's.
d) Write a few sentences comparing the two sluggers.
e) Create a side-by-side stem-and-leaf display comparing the careers of the two players.
f) What aspects of the distributions are apparent in the stem-and-leaf displays that did not clearly show in the boxplots?

25. Be quick! Avoiding an accident when driving can depend on reaction time. That time, measured from the moment the driver first sees the danger until he or she steps on the brake pedal, is thought to follow a Normal model with a mean of 1.5 seconds and a standard deviation of 0.18 seconds.
a) Use the 68–95–99.7 Rule to draw the Normal model.
b) Write a few sentences describing driver reaction times.
c) What percent of drivers have a reaction time less than 1.25 seconds?
d) What percent of drivers have reaction times between 1.6 and 1.8 seconds?
e) What is the interquartile range of reaction times?
f) Describe the reaction times of the slowest 1/3 of all drivers.

26. Music and memory. Is it a good idea to listen to music when studying for a big test? In a study conducted by some Statistics students, 62 people were randomly assigned to listen to rap music, Mozart, or no music

while attempting to memorize objects pictured on a page. They were then asked to list all the objects they could remember. Here are the 5-number summaries for each group:

	n	Min	Q1	Median	Q3	Max
Rap	29	5	8	10	12	25
Mozart	20	4	7	10	12	27
None	13	8	9.5	13	17	24

a) Describe the W's for these data: *Who, What, Where, Why, When, How.*
b) Name the variables and classify each as categorical or quantitative.
c) Create parallel boxplots as best you can from these summary statistics to display these results.
d) Write a few sentences comparing the performances of the three groups.

T **27. Mail.** Here are the number of pieces of mail received at a school office for 36 days.

123	70	90	151	115	97
80	78	72	100	128	130
52	103	138	66	135	76
112	92	93	143	100	88
118	118	106	110	75	60
95	131	59	115	105	85

a) Plot these data.
b) Find appropriate summary statistics.
c) Write a brief description of the school's mail deliveries.
d) What percent of the days actually lie within one standard deviation of the mean? Comment.

T **28. Birth order.** Is your birth order related to your choice of major? A Statistics professor at a large university polled his students to find out what their majors were and what position they held in the family birth order. The results are summarized in the table.

a) What percent of these students are oldest or only children?
b) What percent of Humanities majors are oldest children?
c) What percent of oldest children are Humanities students?
d) What percent of the students are oldest children majoring in the Humanities?

	Birth Order*				
Major	**1**	**2**	**3**	**4+**	Total
Math/Science	34	14	6	3	**57**
Agriculture	52	27	5	9	**93**
Humanities	15	17	8	3	**43**
Other	12	11	1	6	**30**
Total	**113**	**69**	**20**	**21**	**223**

* 1 = oldest or only child

29. Herbal medicine. Researchers for the Herbal Medicine Council collected information on people's experiences with a new herbal remedy for colds. They went to a store that sold natural health products. There they asked 100 customers whether they had taken the cold remedy and, if so, to rate its effectiveness (on a scale from 1 to 10) in curing their symptoms. The Council concluded that this product was highly effective in treating the common cold.
a) Identify the W's of these data.
b) Identify the variables, classify each as categorical or quantitative, and specify units if relevant.
c) Are you confident about the Council's conclusion? Explain.

T **30. Birth order revisited.** Consider again the data on birth order and college majors in Exercise 28.
a) What is the marginal distribution of majors?
b) What is the conditional distribution of majors for the oldest children?
c) What is the conditional distribution of majors for the children born second?
d) Do you think that college major appears to be independent of birth order? Explain.

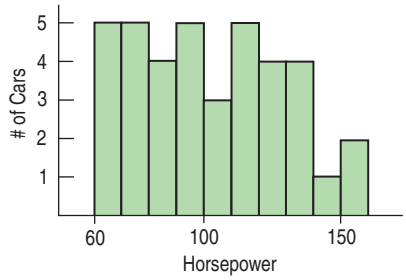
31. Engines. One measure of the size of an automobile engine is its "displacement," the total volume (in liters or cubic inches) of its cylinders. Summary statistics for several models of new cars are shown. These displacements were measured in cubic inches.

Summary of Displacement	
Count	38
Mean	177.29
Median	148.5
StdDev	88.88
Range	275
25th %tile	105
75th %tile	231

a) How many cars were measured?
b) Why might the mean be so much larger than the median?
c) Describe the center and spread of this distribution with appropriate statistics.
d) Your neighbor is bragging about the 227-cubic-inch engine he bought in his new car. Is that engine unusually large? Explain.
e) Are there any engines in this data set that you would consider to be outliers? Explain.
f) Is it reasonable to expect that about 68% of car engines measure between 88 and 266 cubic inches? (That's 177.289 ± 88.8767.) Explain.
g) We can convert all the data from cubic inches to cubic centimeters (cc) by multiplying by 16.4. For example, a 200-cubic-inch engine has a displacement of 3280 cc. How would such a conversion affect each of the summary statistics?

32. Engines, again. Horsepower is another measure commonly used to describe auto engines. Here are the summary statistics and histogram displaying horsepowers of the same group of 38 cars discussed in Exercise 31.

Summary of Horsepower	
Count	38
Mean	101.7
Median	100
StdDev	26.4
Range	90
25th %tile	78
75th %tile	125



a) Describe the shape, center, and spread of this distribution.
b) What is the interquartile range?
c) Are any of these engines outliers in terms of horsepower? Explain.
d) Do you think the 68–95–99.7 Rule applies to the horsepower of auto engines? Explain.
e) From the histogram, make a rough estimate of the percentage of these engines whose horsepower is within one standard deviation of the mean.
f) A fuel additive boasts in its advertising that it can "add 10 horsepower to any car." Assuming that is true, what would happen to each of these summary statistics if this additive were used in all the cars?

33. Age and party 2007. The Pew Research Center conducts surveys regularly asking respondents which political party they identify with. Among their results is the following table relating preferred political party and age. (http://people-press.org/reports/)

		Party			
		Republican	Democrat	Others	Total
Age	18–29	2636	2738	4765	10139
	30–49	6871	6442	8160	21473
	50–64	3896	4286	4806	12988
	65+	3131	3718	2934	9784
	Total	16535	17183	20666	54384

a) What percent of people surveyed were Republicans?
b) Do you think this might be a reasonable estimate of the percentage of all voters who are Republicans? Explain.
c) What percent of people surveyed were under 30 or over 65?
d) What percent of people were classified as "Other" and under the age of 30?
e) What percent of the people classified as "Other" were under 30?
f) What percent of people under 30 were classified as "Other"?

34. Pay. According to the *2006 National Occupational Employment and Wage Estimates for Management Occupations*, the mean hourly wage for Chief Executives was $69.52 and the median hourly wage was "over $70.00." By contrast, for General and Operations Managers, the mean hourly wage was $47.73 and the median was $40.97. Are these wage distributions likely to be symmetric, skewed left, or skewed right? Explain.

35. Age and party II. Consider again the Pew Research Center results on age and political party in Exercise 33.
a) What is the marginal distribution of party affiliation?
b) Create segmented bar graphs displaying the conditional distribution of party affiliation for each age group.
c) Summarize these poll results in a few sentences that might appear in a newspaper article about party affiliation in the United States.
d) Do you think party affiliation is independent of the voter's age? Explain.

T **36. Bike safety 2003.** The Bicycle Helmet Safety Institute website includes a report on the number of bicycle fatalities per year in the United States. The table below shows the counts for the years 1994–2003.

Year	Bicycle fatalities
1994	796
1995	828
1996	761
1997	811
1998	757
1999	750
2000	689
2001	729
2002	663
2003	619

a) What are the W's for these data?
b) Display the data in a stem-and-leaf display.
c) Display the data in a timeplot.
d) What is apparent in the stem-and-leaf display that is hard to see in the timeplot?
e) What is apparent in the timeplot that is hard to see in the stem-and-leaf display?
f) Write a few sentences about bicycle fatalities in the United States.

37. Some assembly required. A company that markets build-it-yourself furniture sells a computer desk that is advertised with the claim "less than an hour to assemble." However, through postpurchase surveys the company has learned that only 25% of its customers succeeded in building the desk in under an hour. The mean time was 1.29 hours. The company assumes that consumer assembly time follows a Normal model.

a) Find the standard deviation of the assembly time model.
b) One way the company could solve this problem would be to change the advertising claim. What assembly time should the company quote in order that 60% of customers succeed in finishing the desk by then?
c) Wishing to maintain the "less than an hour" claim, the company hopes that revising the instructions and labeling the parts more clearly can improve the 1-hour success rate to 60%. If the standard deviation stays the same, what new lower mean time does the company need to achieve?
d) Months later, another postpurchase survey shows that new instructions and part labeling did lower the mean assembly time, but only to 55 minutes. Nonetheless, the company did achieve the 60%-in-an-hour goal, too. How was that possible?

T **38. Profits.** Here is a stem-and-leaf display showing profits as a percent of sales for 29 of the *Forbes* 500 largest U.S. corporations. The stems are split; each stem represents a span of 5%, from a loss of 9% to a profit of 25%.

−0	99
−0	1 2 3 4
0	1 1 1 1 2 3 4 4 4
0	5 5 5 5 6 7 9
1	0 0 1 1 3
1	
2	2
2	5

Profits (% of sales)
(−0|3 means a loss of 3%)

a) Find the 5-number summary.
b) Draw a boxplot for these data.
c) Find the mean and standard deviation.
d) Describe the distribution of profits for these corporations.