CHAPTER

5

Understanding and Comparing Distributions

WHO	Days during 1989
WHAT	Average daily wind speed (mph), Average barometric pressure (mb), Average daily temperature (deg Celsius)
WHEN	1989
WHERE	Hopkins Forest, in Western Massachusetts
WHY	Long-term observations to study ecology and climate

The Hopkins Memorial Forest is a 2500-acre reserve in Massachusetts, New York, and Vermont managed by the Williams College Center for Environmental Studies (CES). As part of their mission, CES monitors forest resources and conditions over the long term. They post daily measurements at their Web site.[1] You can go there, download, and analyze data for any range of days. We'll focus for now on 1989. As we'll see, some interesting things happened that year.

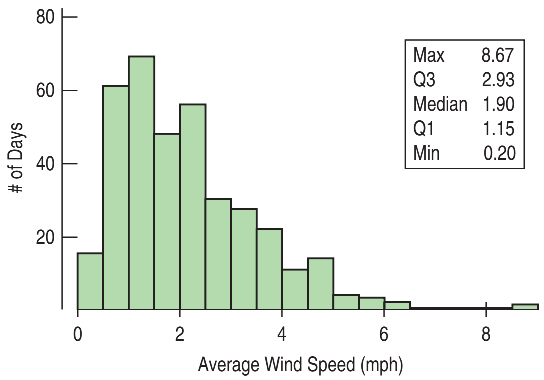
One of the variables measured in the forest is wind speed. Three remote anemometers generate far too much data to report, so, as summaries, you'll find the minimum, maximum, and average wind speed (in mph) for each day.

Wind is caused as air flows from areas of high pressure to areas of low pressure. Centers of low pressure often accompany storms, so both high winds and low pressure are associated with some of the fiercest storms. Wind speeds can vary greatly during a day and from day to day, but if we step back a bit farther, we can see patterns. By modeling these patterns, we can understand things about *Average Wind Speed* that we may not have known.

In Chapter 3 we looked at the association between two categorical variables using contingency tables and displays. Here we'll explore different ways of examining the relationship between two variables when one is quantitative, and the other is categorical and indicates groups to compare. We are given wind speed averages for each day of 1989. But we can collect the days together into different size groups and compare the wind speeds among them. If we consider *Time* as a categorical variable in this way, we'll gain enormous flexibility for our analysis and for our understanding. We'll discover new insights as we change the granularity of the grouping variable—from viewing the whole year's data at one glance, to comparing seasons, to looking for patterns across months, and, finally, to looking at the data day by day.

[1] www.williams.edu/CES/hopkins.htm

The Big Picture



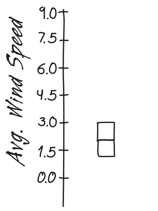
Let's start with the "big picture." Here's a histogram and 5-number summary of the *Average Wind Speed* for every day in 1989. Because of the skewness, we'll report the median and IQR. We can see that the distribution of *Average Wind Speed* is unimodal and skewed to the right. Median daily wind speed is about 1.90 mph, and on half of the days, the average wind speed is between 1.15 and 2.93 mph. We also see a rather windy 8.67-mph day. Was that unusually windy or just the windiest day of the year? To answer that, we'll need to work with the summaries a bit more.

FIGURE 5.1

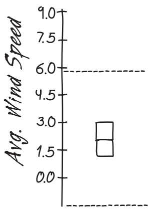
A histogram of daily Average Wind Speed *for 1989. It is unimodal and skewed to the right, with a possible high outlier.*

Boxplots and 5-Number Summaries

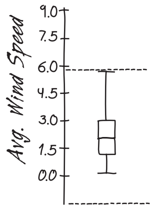
1.



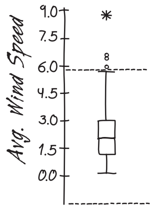
2.



3.



4.



Once we have a 5-number summary of a (quantitative) variable, we can display that information in a **boxplot.** To make a boxplot of the average wind speeds, follow these steps:

1. Draw a single vertical axis spanning the extent of the data.[2] Draw short horizontal lines at the lower and upper quartiles and at the median. Then connect them with vertical lines to form a box. The box can have any width that looks OK.[3]
2. To help us construct the boxplot, we erect "fences" around the main part of the data. We place the upper fence 1.5 IQRs above the upper quartile and the lower fence 1.5 IQRs below the lower quartile. For the wind speed data, we compute

$$Upper\ fence = Q3 + 1.5\ IQR = 2.93 + 1.5 \times 1.78 = 5.60 \text{ mph}$$

and

$$Lower fence = Q1 - 1.5\ IQR = 1.15 - 1.5 \times 1.78 = -1.52 \text{ mph}$$

The fences are just for construction and are not part of the display. We show them here with dotted lines for illustration. You should never include them in your boxplot.
3. We use the fences to grow "whiskers." Draw lines from the ends of the box up and down to *the most extreme data values found within the fences*. If a data value falls outside one of the fences, we do *not* connect it with a whisker.
4. Finally, we add the **outliers** by displaying any data values beyond the fences with special symbols. (We often use a different symbol for "**far outliers**"—data values farther than 3 IQRs from the quartiles.)

What does a boxplot show? The center of a boxplot is (remarkably enough) a box that shows the middle half of the data, between the quartiles. The height of the box is equal to the IQR. If the median is roughly centered between the quartiles, then the middle half of the data is roughly symmetric. If the median is not centered, the distribution is skewed. The whiskers show skewness as well if they are not roughly the same length. Any outliers are displayed individually, both to keep them out of the way for judging skewness and to encourage you to give them special attention. They may be mistakes, or they may be the most interesting cases in your data.

A S **Boxplots.** Watch a boxplot under construction.

TI-*nspire*

Boxplots and dotplots. Drag data points around to explore what a boxplot shows (and doesn't).

[2] The axis could also run horizontally.

[3] Some computer programs draw wider boxes for larger data sets. That can be useful when comparing groups.

The prominent statistician John W. Tukey, the originator of the boxplot, was asked by one of the authors why the outlier nomination rule cut at 1.5 IQRs beyond each quartile. He answered that the reason was that 1 IQR would be too small and 2 IQRs would be too large. That works for us.

For the Hopkins Forest data, the central box contains each day whose *Average Wind Speed* is between 1.15 and 2.93 miles per hour (see Figure 5.2). From the shape of the box, it looks like the central part of the distribution of wind speeds is roughly symmetric, but the longer upper whisker indicates that the distribution stretches out at the upper end. We also see a few very windy days. Boxplots are particularly good at pointing out outliers. These extraordinarily windy days may deserve more attention. We'll give them that extra attention shortly.

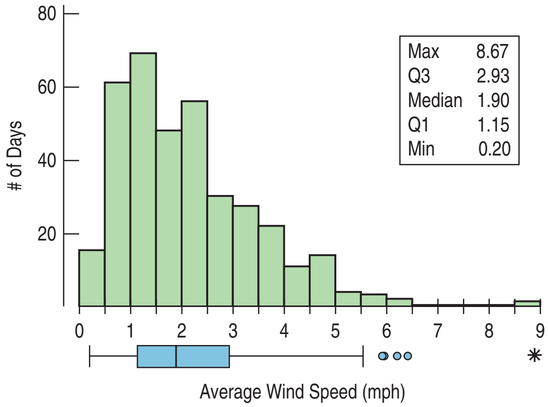


FIGURE 5.2

By turning the boxplot and putting it on the same scale as the histogram, we can compare both displays of the daily wind speeds and see how each represents the distribution.

A S ***Activity:* Playing with Summaries.** See how different summary measures behave as you place and drag values, and see how sensitive some statistics are to individual data values.

Comparing Groups with Histograms

TI-*nspire*

Histograms and boxplots. See that the shape of a distribution is not always evident in a boxplot.

It is almost always more interesting to compare groups. Is it windier in the winter or the summer? Are any months particularly windy? Are weekends a special problem? Let's split the year into two groups: April through September (Spring/Summer) and October through March (Fall/Winter). To compare the groups, we create two histograms, being careful to use the same scale. Here are displays of the average daily wind speed for Spring/Summer (on the left) and Fall/Winter (on the right):

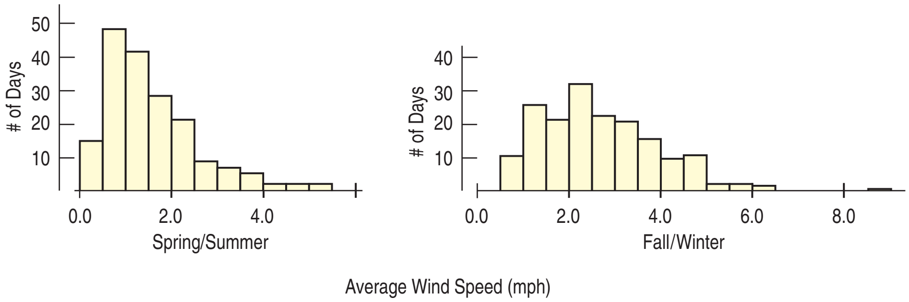


FIGURE 5.3

Histograms of Average Wind Speed *for days in Spring/Summer (left) and Fall/Winter (right) show very different patterns.*

The shapes, centers, and spreads of these two distributions are strikingly different. During spring and summer (histogram on the left), the distribution is skewed to the right. A typical day during these warmer months has an average wind speed of only 1 to 2 mph, and few have average speeds above 3 mph. In the colder months (histogram on the right), however, the shape is less strongly skewed and more spread out. The typical wind speed is higher, and days with average wind speeds above 3 mph are not unusual. There are several noticeable high values.

Summaries for *Average Wind Speed* by Season

Group	Mean	StdDev	Median	IQR
Fall/Winter	2.71	1.36	2.47	1.87
Spring/Summer	1.56	1.01	1.34	1.32

FOR EXAMPLE **Comparing groups with stem-and-leaf displays**

In 2004 the infant death rate in the United States was 6.8 deaths per 1000 live births. The Kaiser Family Foundation collected data from all 50 states and the District of Columbia, allowing us to look at different regions of the country. Since there are only 51 data values, a back-to-back stem-and-leaf plot is an effective display. Here's one comparing infant death rates in the Northeast and Midwest to those in the South and West. In this display the stems run down the middle of the plot, with the leaves for the two regions to the left or right. Be careful when you read the values on the left: 4 | 11 | means a rate of 11.4 deaths per 1000 live birth for one of the southern or western states.

Infant Death Rates (by state) 2004

South and West		North and Midwest
4	11	
3 0	10	
0 0	9	
0 4 1 6 9 5 8	8	1 0
0 5 0 3	7	5 8 0 7 4 1
4 1 0 4 9 1 1 6 4	6	3 1 5 4 4
6 3 6 2	5	8 4 0 6
	4	8 8 9 7
	3	

(4 |11| means 11.4 deaths per 1000 live births)

Question: How do infant death rates compare for these regions?

In general, infant death rates were generally higher for states in the South and West than in the Northeast and Midwest. The distribution for the northeastern and midwestern states is roughly uniform, varying from a low of 4.8 to a high of 8.1 deaths per 1000 live births. Ten southern and western states had higher infant death rates than any in the Northeast or Midwest, with one state over 11. Rates varied more widely in the South and West, where the distribution is skewed to the right and possibly bimodal. We should investigate further to see which states represent the cluster of high death rates.

Comparing Groups with Boxplots

A S ***Video:*** **Can Diet Prolong Life?** Here's a subject that's been in the news: Can you live longer by eating less? (Or would it just seem longer?) Look at the data in subsequent activities, and you'll find that you can learn a lot by comparing two groups with boxplots.

Are some months windier than others? Even residents may not have a good idea of which parts of the year are the most windy. (Do you know for your hometown?) We're not interested just in the centers, but also in the spreads. Are wind speeds equally variable from month to month, or do some months show more variation?

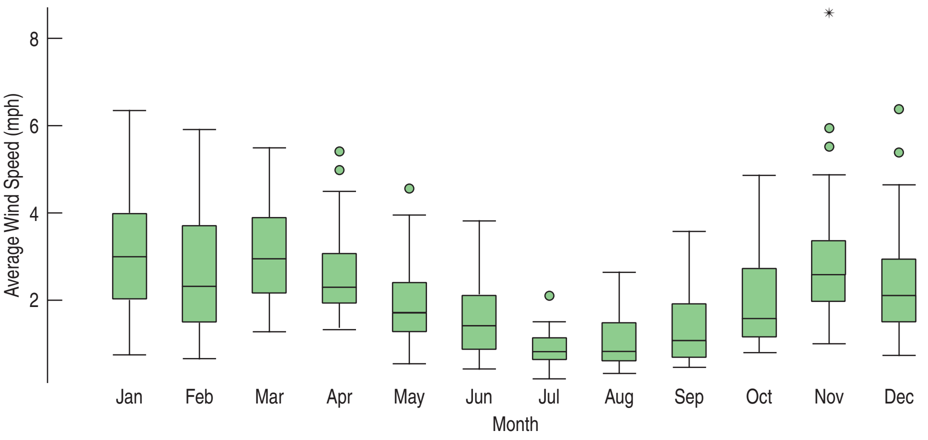
Earlier, we compared histograms of the wind speeds for two halves of the year. To look for seasonal trends, though, we'll group the daily observations by month. Histograms or stem-and-leaf displays are a fine way to look at one distribution or two. But it would be hard to see patterns by comparing 12 histograms. Boxplots offer an ideal balance of information and simplicity, hiding the details while displaying the overall summary information. So we often plot them side by side for groups or categories we wish to compare.

By placing boxplots side by side, we can easily see which groups have higher medians, which have the greater IQRs, where the central 50% of the data is located in each group, and which have the greater overall range. And, when the boxes are in an order, we can get a general idea of patterns in both the centers and the spreads. Equally important, we can see past any outliers in making these comparisons because they've been displayed separately.

Here are boxplots of the *Average Daily Wind Speed* by month:

FIGURE 5.4

Boxplots of the average daily wind speed for each month show seasonal patterns in both the centers and spreads.

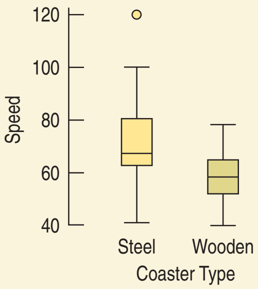


Here we see that wind speeds tend to decrease in the summer. The months in which the winds are both strongest and most variable are November through March. And there was one remarkably windy day in November.

When we looked at a boxplot of wind speeds for the entire year, there were only 5 outliers. Now, when we group the days by *Month,* the boxplots display more days as outliers and call out one in November as a far outlier. The boxplots show different outliers than before because some days that seemed ordinary when placed against the entire year's data looked like outliers for the month that they're in. That windy day in July certainly wouldn't stand out in November or December, but for July, it was remarkable.

FOR EXAMPLE Comparing distributions

Roller coasters[4] are a thrill ride in many amusement parks worldwide. And thrill seekers want a coaster that goes fast. There are two main types of roller coasters: those with wooden tracks and those with steel tracks. Do they typically run at different speeds? Here are boxplots:



Question: Compare the speeds of wood and steel roller coasters.

Overall, wooden-track roller coasters are slower than steel-track coasters. In fact, the fastest half of the steel coasters are faster than three quarters of the wooden coasters. Although the IQRs of the two groups are similar, the range of speeds among steel coasters is larger than the range for wooden coasters. The distribution of speeds of wooden coasters appears to be roughly symmetric, but the speeds of the steel coasters are skewed to the right, and there is a high outlier at 120 mph. We should look into why that steel coaster is so fast.

STEP-BY-STEP EXAMPLE Comparing Groups

Of course, we can compare groups even when they are not in any particular order. Most scientific studies compare two or more groups. It is almost always a good idea to start an analysis of data from such studies by comparing boxplots for the groups. Here's an example:

For her class project, a student compared the efficiency of various coffee containers. For her study, she decided to try 4 different containers and to test each of them 8 different times. Each time, she heated water to 180°F, poured it into a container, and sealed it. (We'll learn the details of how to set up experiments in Chapter 13.) After 30 minutes, she measured the temperature again and recorded the difference in temperature. Because these are temperature differences, smaller differences mean that the liquid stayed hot—just what we would want in a coffee mug.

Question: What can we say about the effectiveness of these four mugs?

[4] See the Roller Coaster Data Base at www.rcdb.com.

THINK

Plan State what you want to find out.

Variables Identify the *variables* and report the W's.

Be sure to check the appropriate condition.

I want to compare the effectiveness of the different mugs in maintaining temperature. I have 8 measurements of *Temperature Change* for each of the mugs.

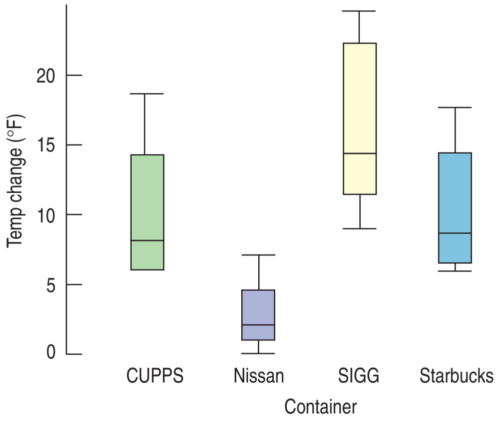
✓ **Quantitative Data Condition:** The *Temperature Changes* are quantitative, with units of °F. Boxplots are appropriate displays for comparing the groups. Numerical summaries of each group are appropriate as well.

SHOW

Mechanics Report the 5-number summaries of the four groups. Including the IQR is a good idea as well.

	Min	Q1	Median	Q3	Max	IQR
CUPPS	6°F	6	8.25	14.25	18.50	8.25
Nissan	0	1	2	4.50	7	3.50
SIGG	9	11.50	14.25	21.75	24.50	10.25
Starbucks	6	6.50	8.50	14.25	17.50	7.75

Make a picture. Because we want to compare the distributions for four groups, boxplots are an appropriate choice.



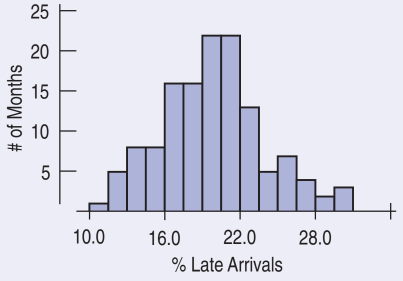
TELL

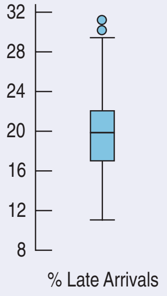
Conclusion Interpret what the boxplots and summaries say about the ability of these mugs to retain heat. Compare the shapes, centers, and spreads, and note any outliers.

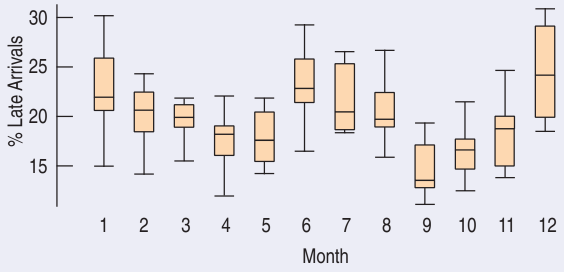
The individual distributions of temperature changes are all slightly skewed to the high end. The Nissan cup does the best job of keeping liquids hot, with a median loss of only 2°F, and the SIGG cup does the worst, typically losing 14°F. The difference is large enough to be important: A coffee drinker would be likely to notice a 14° drop in temperature. And the mugs are clearly different: 75% of the Nissan tests showed less heat loss than any of the other mugs in the study. The IQR of results for the Nissan cup is also the smallest of these test cups, indicating that it is a consistent performer.

JUST CHECKING

The Bureau of Transportation Statistics of the U.S. Department of Transportation collects and publishes statistics on airline travel (www.transtats.bts.gov). Here are three displays of the % of flights arriving late each month from 1995 through 2005:







1. Describe what the histogram says about late arrivals.
2. What does the boxplot of late arrivals suggest that you can't see in the histogram?
3. Describe the patterns shown in the boxplots by month. At what time of year are flights least likely to be late? Can you suggest reasons for this pattern?

TI Tips

Comparing groups with boxplots

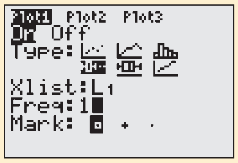
In the last chapter we looked at the performances of fourth-grade students on an agility test. Now let's make comparative boxplots for the boys' scores and the girls' scores:

Boys: 22, 17, 18, 29, 22, 22, 23, 24, 23, 17, 21

Girls: 25, 20, 12, 19, 28, 24, 22, 21, 25, 26, 25, 16, 27, 22

Enter these data in L1 (*Boys*) and L2 (*Girls*).

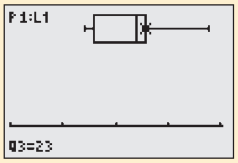
Set up STATPLOT's Plot1 to make a boxplot of the boys' data:



Plot1 Plot2 Plot3
On Off
Type:
Xlist:L1
Freq:1
Mark: □ + ·

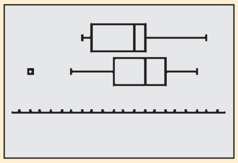
- Turn the plot On;
- Choose the first boxplot icon (you want your plot to indicate outliers);
- Specify Xlist:L1 and Freq:1, and select the Mark you want the calculator to use for displaying any outliers.

Use ZoomStat to display the boxplot for *Boys*. You can now TRACE to see the statistics in the five-number summary. Try it!



P1:L1
Q3=23

As you did for the boys, set up Plot2 to display the girls' data. This time when you use ZoomStat with both plots turned on, the display shows the parallel boxplots. See the outlier?



This is a great opportunity to practice your "Tell" skills. How do these fourth graders compare in terms of agility?

Outliers

When we looked at boxplots for the *Average Wind Speed* by *Month*, we noticed that several days stood out as possible outliers and that one very windy day in November seemed truly remarkable. What should we do with such outliers?

Cases that stand out from the rest of the data almost always deserve our attention. An outlier is a value that doesn't fit with the rest of the data, but exactly how different it should be to be treated specially is a judgment call. Boxplots provide a rule of thumb to highlight these unusual points, but that rule doesn't tell you what to do with them.

So, what *should* we do with outliers? The first thing to do is to try to understand them in the context of the data. A good place to start is with a histogram. Histograms show us more detail about a distribution than a boxplot can, so they give us a better idea of how the outlier fits (or doesn't fit) in with the rest of the data.

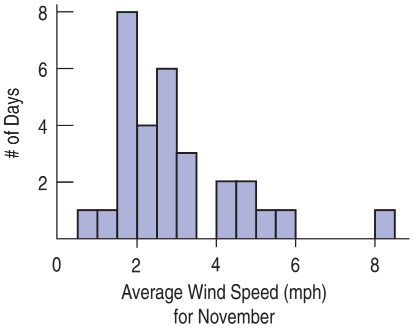


FIGURE 5.5

The Average Wind Speed *in November is slightly skewed with a high outlier.*

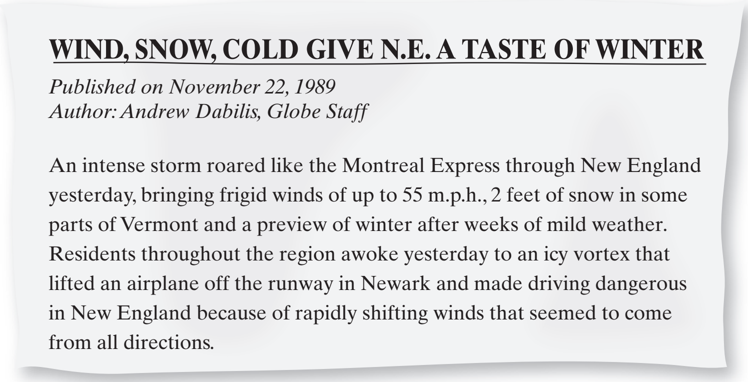
A histogram of the *Average Wind Speed* in November shows a slightly skewed main body of data and that very windy day clearly set apart from the other days. When considering whether a case is an outlier, we often look at the gap between that case and the rest of the data. A large gap suggests that the case really is quite different. But a case that just happens to be the largest or smallest value at the end of a possibly stretched-out tail may be best thought of as just . . . the largest or smallest value. After all, *some* case has to be the largest or smallest.

Some outliers are simply unbelievable. If a class survey includes a student who claims to be 170 inches tall (about 14 feet, or 4.3 meters), you can be pretty sure that's an error.

Once you've identified likely outliers, you should always investigate them. Some outliers are just errors. A decimal point may have been misplaced, digits transposed, or digits repeated or omitted. The units may be wrong. (Was that outlying height reported in centimeters rather than in inches [170 cm = 65 in.]?) Or a number may just have been transcribed incorrectly, perhaps copying an adjacent value on the original data sheet. If you can identify the correct value, then you should certainly fix it. One important reason to look into outliers is to correct errors in your data.

Many outliers are not wrong; they're just different. Such cases often repay the effort to understand them. You can learn more from the extraordinary cases than from summaries of the overall data set.

What about that windy November day? Was it really that windy, or could there have been a problem with the anemometers? A quick Internet search for weather on November 21, 1989, finds that there was a severe storm:



WIND, SNOW, COLD GIVE N.E. A TASTE OF WINTER

Published on November 22, 1989
Author: Andrew Dabilis, Globe Staff

An intense storm roared like the Montreal Express through New England yesterday, bringing frigid winds of up to 55 m.p.h., 2 feet of snow in some parts of Vermont and a preview of winter after weeks of mild weather. Residents throughout the region awoke yesterday to an icy vortex that lifted an airplane off the runway in Newark and made driving dangerous in New England because of rapidly shifting winds that seemed to come from all directions.

When we have outliers, we need to decide what to *Tell* about the data. If we can correct an error, we'll just summarize the corrected data (and note the correction). But if we see no way to correct an outlying value, or if we confirm that it is correct, our best path is to report summaries and analyses with *and* without the outlier. In this way a reader can judge for him- or herself what influence the outlier has and decide what to think about the data.

There are two things we should *never* do with outliers. The first is to silently leave an outlier in place and proceed as if nothing were unusual. Analyses of data with outliers are very likely to be influenced by those outliers—sometimes to a large and misleading degree. The other is to drop an outlier from the analysis without comment just because it's unusual. If you want to exclude an outlier, you must discuss your decision and, to the extent you can, justify your decision.

A S ***Case Study:* Are passengers or drivers safer in a crash?** Practice the skills of this chapter by comparing these two groups.

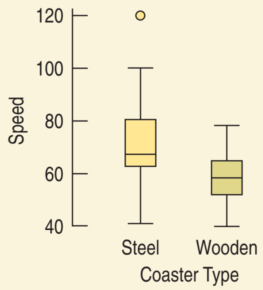
FOR EXAMPLE Checking out the outliers

Recap: We've looked at the speeds of roller coasters and found a difference between steel- and wooden-track coasters. We also noticed an extraordinary value.

Question: The fastest coaster in this collection turns out to be the "Top Thrill Dragster" at Cedar Point amusement park. What might make this roller coaster unusual? You'll have to do some research, but that's often what happens with outliers.

The Top Thrill Dragster is easy to find in an Internet search. We learn that it is a "hydraulic launch" coaster. That is, it doesn't get its remarkable speed just from gravity, but rather from a kick-start by a hydraulic piston. That could make it different from the other roller coasters.

(You might also discover that it is no longer the fastest roller coaster in the world.)



Timeplots: Order, Please!

The Hopkins Forest wind speeds are reported as daily averages. Previously, we grouped the days into months or seasons, but we could look at the wind speed values day by day. Whenever we have data measured over time, it is a good idea to look for patterns by plotting the data in time order. Here are the daily average wind speeds plotted over time:

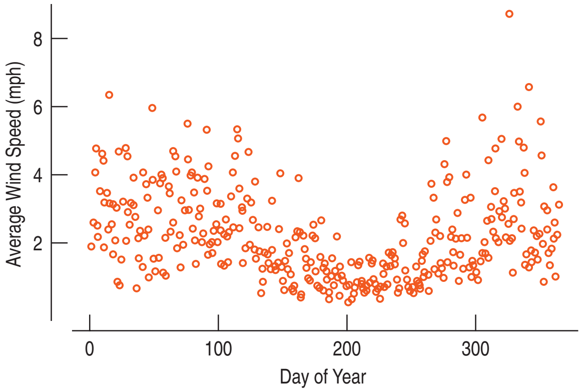


FIGURE 5.6

A timeplot of Average Wind Speed *shows the overall pattern and changes in variation.*

A display of values against time is sometimes called a **timeplot**. This timeplot reflects the pattern that we saw when we plotted the wind speeds by month. But without the arbitrary divisions between months, we can see a calm period during the summer, starting around day 200 (the middle of July), when the wind is relatively mild and doesn't vary greatly from day to day. We can also see that the wind becomes both more variable and stronger during the early and late parts of the year.

Looking into the Future

It is always tempting to try to extend what we see in a timeplot into the future. Sometimes that makes sense. Most likely, the Hopkins Forest climate follows regular seasonal patterns. It's probably safe to predict a less windy June next year and a windier November. But we certainly wouldn't predict another storm on November 21.

Other patterns are riskier to extend into the future. If a stock has been rising, will it continue to go up? No stock has ever increased in value indefinitely, and no stock analyst has consistently been able to forecast when a stock's value will turn around. Stock prices, unemployment rates, and other economic, social, or psychological concepts are much harder to predict than physical quantities. The path a ball will follow when thrown from a certain height at a given speed and direction is well understood. The path interest rates will take is much less clear. Unless we have strong (nonstatistical) reasons for doing otherwise, we should resist the temptation to think that any trend we see will continue, even into the near future.

Statistical models often tempt those who use them to think beyond the data. We'll pay close attention later in this book to understanding when, how, and how much we can justify doing that.

Re-expressing Data: A First Look

Re-expressing to Improve Symmetry

When the data are skewed, it can be hard to summarize them simply with a center and spread, and hard to decide whether the most extreme values are outliers or just part of the stretched-out tail. How can we say anything useful about such data? The secret is to *re-express* the data by applying a simple function to each value.

Many relationships and "laws" in the sciences and social sciences include functions such as logarithms, square roots, and reciprocals. Similar relationships often show up in data. Here's a simple example:

In 1980 large companies' chief executive officers (CEOs) made, on average, about 42 times what workers earned. In the next two decades, CEO compensation soared when compared to the average worker. By 2000 that multiple had jumped[5]

[5] *Sources:* United for a Fair Economy, *Business Week* annual CEO pay surveys, Bureau of Labor Statistics, "Average Weekly Earnings of Production Workers, Total Private Sector." Series ID: EEU00500004.

to 525. What does the distribution of the compensation of Fortune 500 companies' CEOs look like? Here's a histogram and boxplot for 2005 compensation:

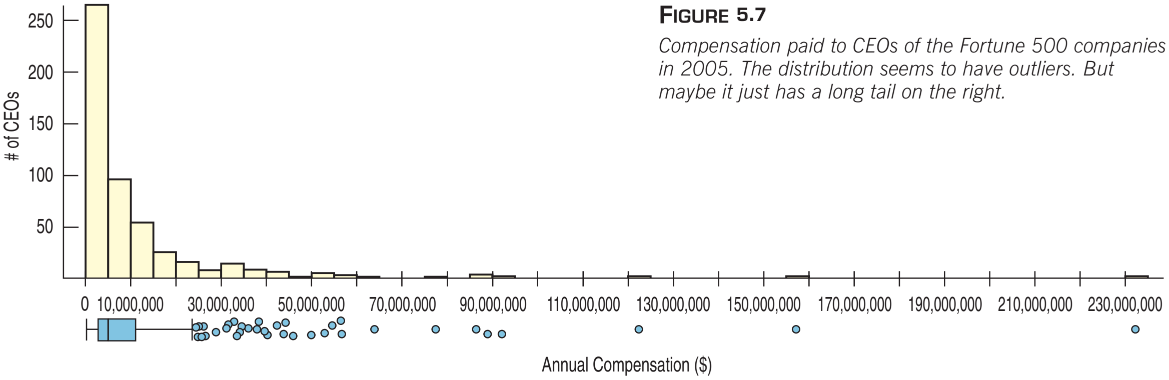


Figure 5.7

Compensation paid to CEOs of the Fortune 500 companies in 2005. The distribution seems to have outliers. But maybe it just has a long tail on the right.

We have 500 CEOs and about 48 possible histogram bins, most of which are empty—but don't miss the tiny bars straggling out to the right. The boxplot indicates that some CEOs received extraordinarily high compensations, while the majority received relatively "little." But look at the values of the bins. The first bin, with about half the CEOs, covers incomes from $0 to $5,000,000. Imagine receiving a salary survey with these categories:

What is your income?
a) $0 to $5,000,000
b) $5,000,001 to $10,000,000
c) $10,000,001 to $15,000,000
d) More than $15,000,000

The reason that the histogram seems to leave so much of the area blank is that the salaries are spread all along the axis from about $15,000,000 to $240,000,000. After $50,000,000 there are so few for each bin that it's very hard to see the tiny bars. What we *can* see from this histogram and boxplot is that this distribution is highly skewed to the right.

It can be hard to decide what we mean by the "center" of a skewed distribution, so it's hard to pick a typical value to summarize the distribution. What would you say was a typical CEO total compensation? The mean value is $10,307,000, while the median is "only" $4,700,000. Each tells us something different about the data.

One approach is to **re-express,** or **transform,** the data by applying a simple function to make the skewed distribution more symmetric. For example, we could take the square root or logarithm of each compensation value. Taking logs works pretty well for the CEO compensations, as you can see:

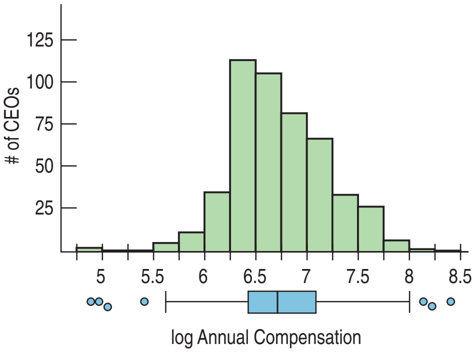


Figure 5.8

The logarithms of 2005 CEO compensations are much more nearly symmetric.

The histogram of the logs of the total CEO compensations is much more nearly symmetric, so we can see that a typical log compensation is between 6, which corresponds to $1,000,000, and 7, corresponding to $10,000,000. And it's easier to talk about a typical value for the logs. The mean log compensation is 6.73, while the median is 6.67. (That's $5,370,317 and $4,677,351, respectively.) Notice that nearly all the values are between 6.0 and 8.0—in other words, between $1,000,000 and $100,000,000 a year, but who's counting?

Against the background of a generally symmetric main body of data, it's easier to decide whether the largest compensations are outliers. In fact, the three most highly compensated CEOs are identified as outliers by the boxplot rule of thumb even after this re-expression. It's perhaps impressive to be an outlier CEO in annual compensation. It's even more impressive to be an outlier in the log scale!

Dealing with logarithms You have probably learned about logs in math courses and seen them in psychology or science classes. In this book, we use them only for making data behave better. Base 10 logs are the easiest to understand, but natural logs are often used as well. (Either one is fine.) You can think of base 10 logs as roughly one less than the number of digits you need to write the number. So 100, which is the smallest number to require 3 digits, has a $\log_{10}$ of 2. And 1000 has a $\log_{10}$ of 3. The $\log_{10}$ of 500 is between 2 and 3, but you'd need a calculator to find that it's approximately 2.7. All salaries of "six figures" have $\log_{10}$ between 5 and 6. Logs are incredibly useful for making skewed data more symmetric. But don't worry—nobody does logs without technology and neither should you. Often, remaking a histogram or other display of the data is as easy as pushing another button.

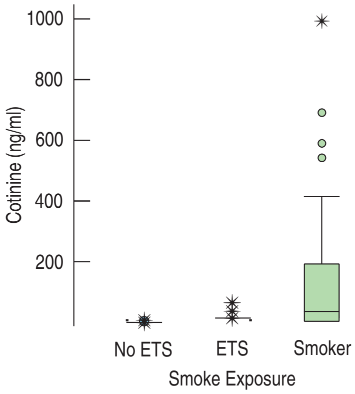


FIGURE 5.9

Cotinine levels (nanograms per milliliter) for three groups with different exposures to tobacco smoke. Can you compare the ETS (exposed to smoke) and No-ETS groups?

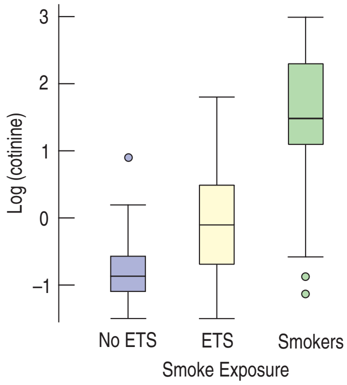


FIGURE 5.10

Blood cotinine levels after taking logs. What a difference a log makes!

Re-expressing to Equalize Spread Across Groups

Researchers measured the concentration (nanograms per milliliter) of cotinine in the blood of three groups of people: nonsmokers who have not been exposed to smoke, nonsmokers who have been exposed to smoke (ETS), and smokers. Cotinine is left in the blood when the body metabolizes nicotine, so this measure gives a direct measurement of the effect of passive smoke exposure. The boxplots of the cotinine levels of the three groups tell us that the smokers have higher cotinine levels, but if we want to compare the levels of the passive smokers to those of the nonsmokers, we're in trouble, because on this scale, the cotinine levels for both nonsmoking groups are too low to be seen.

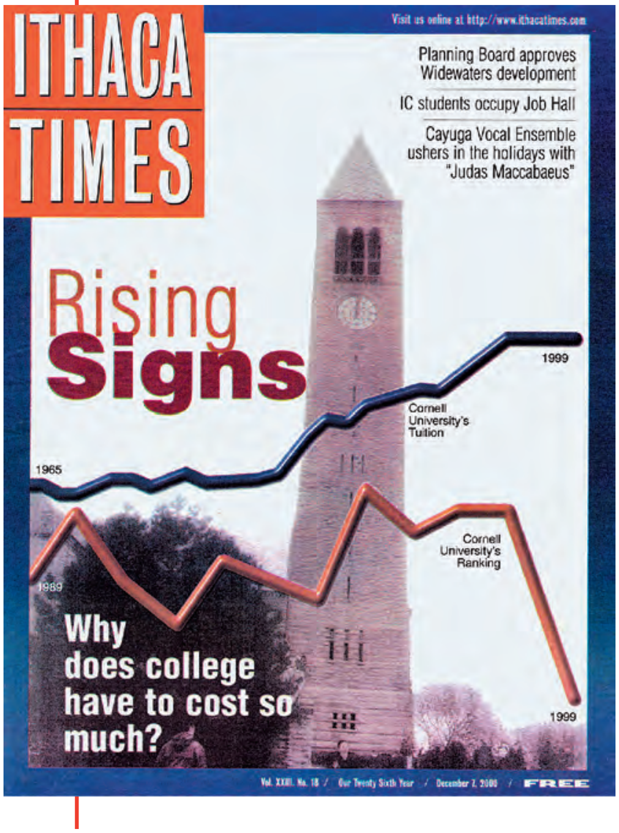
Re-expressing can help alleviate the problem of comparing groups that have very different spreads. For measurements like the cotinine data, whose values can't be negative and whose distributions are skewed to the high end, a good first guess at a re-expression is the logarithm.

After taking logs, we can compare the groups and see that the nonsmokers exposed to environmental smoke (the ETS group) do show increased levels of (log) cotinine, although not the high levels found in the blood of smokers.

Notice that the same re-expression has also improved the symmetry of the cotinine distribution for smokers and pulled in most of the apparent outliers in all of the groups. It is not unusual for a re-expression that improves one aspect of data to improve others as well. We'll talk about other ways to re-express data as the need arises throughout the book. We'll explore some common re-expressions more thoroughly in Chapter 10.

WHAT CAN GO WRONG?

- **Avoid inconsistent scales.** Parts of displays should be mutually consistent—no fair changing scales in the middle or plotting two variables on different scales but on the same display. When comparing two groups, be sure to compare them on the same scale.
- **Label clearly.** Variables should be identified clearly and axes labeled so a reader knows what the plot displays.



Visit us online at http://www.ithacatimes.com

ITHACA TIMES

Planning Board approves Widewaters development

IC students occupy Job Hall

Cayuga Vocal Ensemble ushers in the holidays with "Judas Maccabaeus"

Rising Signs

Why does college have to cost so much?

Vol. XXIII. No. 18 / Our Twenty Sixth Year / December 7, 2000 / FREE

Here's a remarkable example of a plot gone wrong. It illustrated a news story about rising college costs. It uses timeplots, but it gives a misleading impression. First think about the story you're being told by this display. Then try to figure out what has gone wrong.

What's wrong? Just about everything.

- The horizontal scales are inconsistent. Both lines show trends over time, but exactly for what years? The tuition sequence starts in 1965, but rankings are graphed from 1989. Plotting them on the same (invisible) scale makes it seem that they're for the same years.
- The vertical axis isn't labeled. That hides the fact that it's inconsistent. Does it graph dollars (of tuition) or ranking (of Cornell University)?

This display violates three of the rules. And it's even worse than that: It violates a rule that we didn't even bother to mention.

- The two inconsistent scales for the vertical axis don't point in the same direction! The line for Cornell's rank shows that it has "plummeted" from 15th place to 6th place in academic rank. Most of us think that's an *improvement*, but that's not the message of this graph.

- **Beware of outliers.** If the data have outliers and you can correct them, you should do so. If they are clearly wrong or impossible, you should remove them and report on them. Otherwise, consider summarizing the data both with and without the outliers.

CONNECTIONS

We discussed the value of summarizing a distribution with shape, center, and spread in Chapter 4, and we developed several ways to measure these attributes. Now we've seen the value of comparing distributions for different groups and of looking at patterns in a quantitative variable measured over time. Although it can be interesting to summarize a single variable for a single group, it is almost always more interesting to compare groups and look for patterns across several groups and over time. We'll continue to make comparisons like these throughout the rest of our work.

WHAT HAVE WE LEARNED?

- We've learned the value of comparing groups and looking for patterns among groups and over time.
- We've seen that boxplots are very effective for comparing groups graphically. When we compare groups, we discuss their shape, center, and spreads, and any unusual features.
- We've experienced the value of identifying and investigating outliers. And we've seen that when we group data in different ways, it can allow different cases to emerge as possible outliers.
- We've graphed data that have been measured over time against a time axis and looked for long-term trends.

Terms

Boxplot

81. A boxplot displays the 5-number summary as a central box with whiskers that extend to the non-outlying data values. Boxplots are particularly effective for comparing groups and for displaying outliers.

Outlier

81, 87. Any point more than 1.5 IQR from either end of the box in a boxplot is nominated as an outlier.

Far Outlier

81. If a point is more than 3.0 IQR from either end of the box in a boxplot, it is nominated as a *far outlier*.

Comparing distributions

82. When comparing the distributions of several groups using histograms or stem-and-leaf displays, consider their:

- Shape
- Center
- Spread

Comparing boxplots

83. When comparing groups with boxplots:

- Compare the shapes. Do the boxes look symmetric or skewed? Are there differences between groups?
- Compare the medians. Which group has the higher center? Is there any pattern to the medians?
- Compare the IQRs. Which group is more spread out? Is there any pattern to how the IQRs change?
- Using the IQRs as a background measure of variation, do the medians seem to be different, or do they just vary much as you'd expect from the overall variation?
- Check for possible outliers. Identify them if you can and discuss why they might be unusual. Of course, correct them if you find that they are errors.

Timeplot

88. A timeplot displays data that change over time. Often, successive values are connected with lines to show trends more clearly. Sometimes a smooth curve is added to the plot to help show long-term patterns and trends.

Skills

THINK

- Be able to select a suitable display for comparing groups. Understand that histograms show distributions well, but are difficult to use when comparing more than two or three groups. Boxplots are more effective for comparing several groups, in part because they show much less information about the distribution of each group.
- Understand that how you group data can affect what kinds of patterns and relationships you are likely to see. Know how to select groupings to show the information that is important for your analysis.
- Be aware of the effects of skewness and outliers on measures of center and spread. Know how to select appropriate measures for comparing groups based on their displayed distributions.
- Understand that outliers can emerge at different groupings of data and that, whatever their source, they deserve special attention.
- Recognize when it is appropriate to make a timeplot.

SHOW

- Know how to make side-by-side histograms on comparable scales to compare the distributions of two groups.
- Know how to make side-by-side boxplots to compare the distributions of two or more groups.
- Know how to describe differences among groups in terms of patterns and changes in their center, spread, shape, and unusual values.
- Know how to make a timeplot of data that have been measured over time.

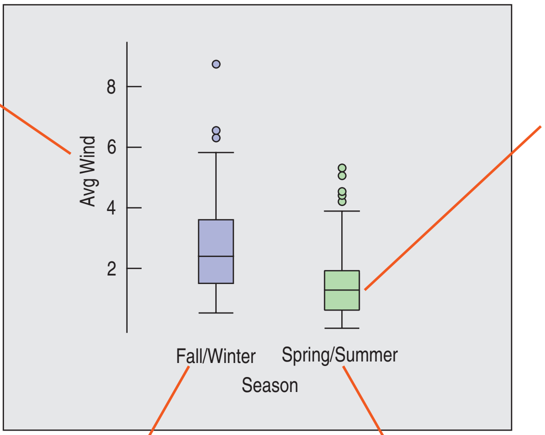
TELL

- Know how to compare the distributions of two or more groups by comparing their shapes, centers, and spreads. Be prepared to explain your choice of measures of center and spread for comparing the groups.
- Be able to describe trends and patterns in the centers and spreads of groups—especially if there is a natural order to the groups, such as a time order.
- Be prepared to discuss patterns in a timeplot in terms of both the general trend of the data and the changes in how spread out the pattern is.
- Be cautious about assuming that trends over time will continue into the future.
- 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 compare the distributions of two or more groups by comparing their shapes, centers, and spreads.
- Know how to describe patterns over time shown in a timeplot.
- Be able to discuss any outliers in the data, noting how they deviate from the overall pattern of the data.

COMPARING DISTRIBUTIONS ON THE COMPUTER

Most programs for displaying and analyzing data can display plots to compare the distributions of different groups. Typically these are boxplots displayed side-by-side.

Side-by-side boxplots should be on the same y-axis scale so they can be compared.



Some programs offer a graphical way to assess how much the medians differ by drawing a band around the median or by "notching" the boxes.

Boxes are typically labeled with a group name. Often they are placed in alphabetical order by group name—not the most useful order.

EXERCISES

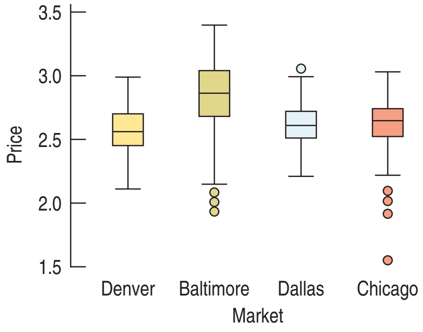
1. **In the news.** Find an article in a newspaper, magazine, or the Internet that compares two or more groups of data.
   a) Does the article discuss the W's?
   b) Is the chosen display appropriate? Explain.
   c) Discuss what the display reveals about the groups.
   d) Does the article accurately describe and interpret the data? Explain.

2. **In the news.** Find an article in a newspaper, magazine, or the Internet that shows a time plot.
   a) Does the article discuss the W's?
   b) Is the timeplot appropriate for the data? Explain.
   c) Discuss what the timeplot reveals about the variable.
   d) Does the article accurately describe and interpret the data? Explain.

3. **Time on the Internet.** Find data on the Internet (or elsewhere) that give results recorded over time. Make an appropriate display and discuss what it shows.

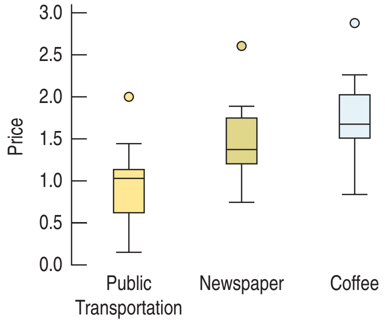
4. **Groups on the Internet.** Find data on the Internet (or elsewhere) for two or more groups. Make appropriate displays to compare the groups, and interpret what you find.

5. **Pizza prices.** A company that sells frozen pizza to stores in four markets in the United States (Denver, Baltimore, Dallas, and Chicago) wants to examine the prices that the stores charge for pizza slices. Here are boxplots comparing data from a sample of stores in each market:



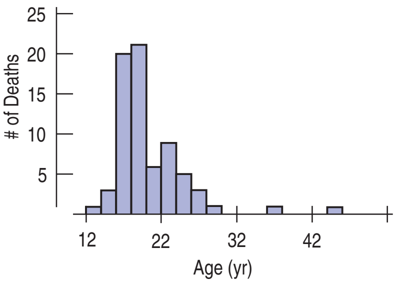
   a) Do prices appear to be the same in the four markets? Explain.
   b) Does the presence of any outliers affect your overall conclusions about prices in the four markets?

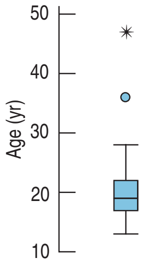
6. **Costs.** To help travelers know what to expect, researchers collected the prices of commodities in 16 cities throughout the world. Here are boxplots comparing the prices of a ride on public transportation, a newspaper, and a cup of coffee in the 16 cities (prices are all in $US).



   a) On average, which commodity is the most expensive?
   b) Is a newspaper always more expensive than a ride on public transportation? Explain.
   c) Does the presence of outliers affect your conclusions in a) or b)?

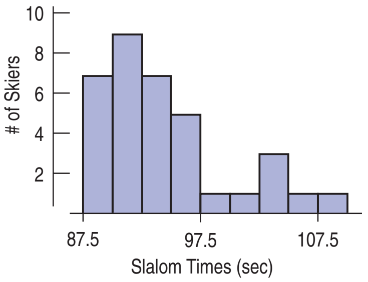
7. **Still rockin'.** Crowd Management Strategies monitors accidents at rock concerts. In their database, they list the names and other variables of victims whose deaths were attributed to "crowd crush" at rock concerts. Here are the histogram and boxplot of the victims' ages for data from 1999 to 2000:

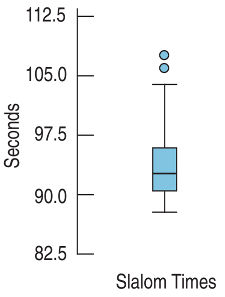




   a) What features of the distribution can you see in both the histogram and the boxplot?
   b) What features of the distribution can you see in the histogram that you could not see in the boxplot?
   c) What summary statistic would you choose to summarize the center of this distribution? Why?
   d) What summary statistic would you choose to summarize the spread of this distribution? Why?

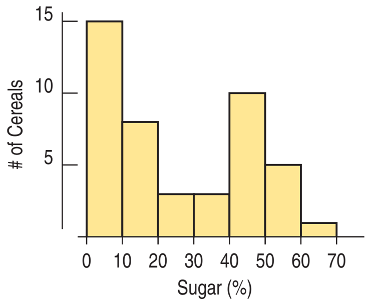
8. **Slalom times.** The Men's Combined skiing event consists of a downhill and a slalom. Here are two displays of the slalom times in the Men's Combined at the 2006 Winter Olympics:

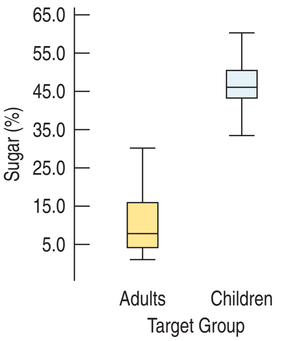




a) What features of the distribution can you see in both the histogram and the boxplot?
b) What features of the distribution can you see in the histogram that you could not see in the boxplot?
c) What summary statistic would you choose to summarize the center of this distribution? Why?
d) What summary statistic would you choose to summarize the spread of this distribution? Why?

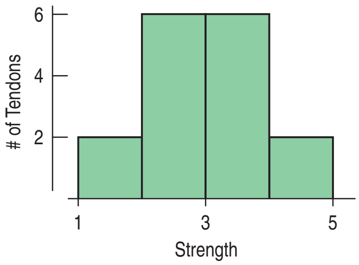
T 9. **Cereals.** Sugar is a major ingredient in many breakfast cereals. The histogram displays the sugar content as a percentage of weight for 49 brands of cereal. The boxplot compares sugar content for adult and children's cereals.

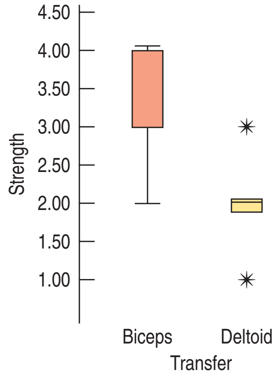




a) What is the range of the sugar contents of these cereals.
b) Describe the shape of the distribution.
c) What aspect of breakfast cereals might account for this shape?
d) Are all children's cereals higher in sugar than adult cereals?
e) Which group of cereals varies more in sugar content? Explain.

T 10. **Tendon transfers.** People with spinal cord injuries may lose function in some, but not all, of their muscles. The ability to push oneself up is particularly important for shifting position when seated and for transferring into and out of wheelchairs. Surgeons compared two operations to restore the ability to push up in children. The histogram shows scores rating pushing strength two years after surgery and boxplots compare results for the two surgical methods. (Mulcahey, Lutz, Kozen, Betz, "Prospective Evaluation of Biceps to Triceps and Deltoid to Triceps for Elbow Extension in Tetraplegia," *Journal of Hand Surgery*, 28, 6, 2003)





a) Describe the shape of this distribution.
b) What is the range of the strength scores?
c) What fact about results of the two procedures is hidden in the histogram?
d) Which method had the higher (better) median score?
e) Was that method always best?
f) Which method produced the most consistent results? Explain.

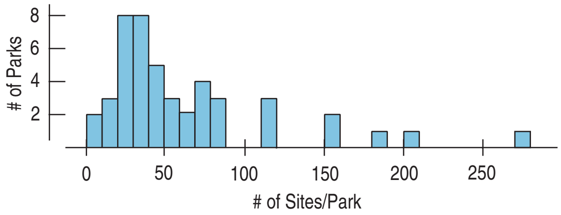
T 11. **Population growth.** Here is a "back-to-back" stem-and-leaf display that shows two data sets at once—one going to the left, one to the right. The display compares the percent change in population for two regions of the United States (based on census figures for 1990 and 2000). The fastest growing states were Nevada at 66% and Arizona at 40%. To show the distributions better, this display breaks each stem into two lines, putting leaves 0–4 on one stem and leaves 5–9 on the other.

NE/MW States		S/W States
	6	6
	6	
	5	
	5	
	4	
	4	0
	3	
	3	001
	2	6
	2	001134
	1	578
2100	1	001134444
99998876655	0	6999
4431	0	1

Population Growth rate
(|6| 6 means 66%)

a) Use the data displayed in the stem-and-leaf display to construct comparative boxplots.
b) Write a few sentences describing the difference in growth rates for the two regions of the United States.

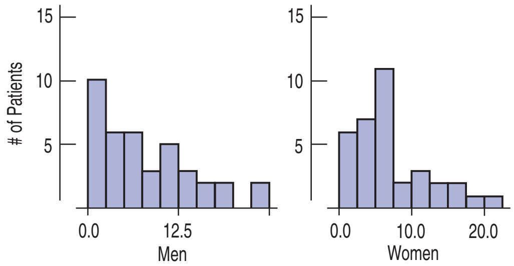
12. **Camp sites.** Shown below are the histogram and summary statistics for the number of camp sites at public parks in Vermont.



Count	46
Mean	62.8 sites
Median	43.5
StdDev	56.2
Min	0
Max	275
Q1	28
Q3	78

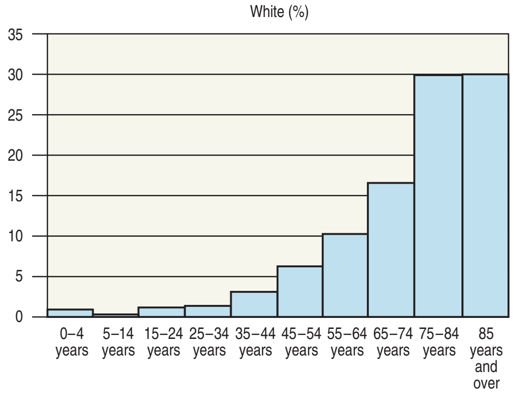
a) Which statistics would you use to identify the center and spread of this distribution? Why?
b) How many parks would you classify as outliers? Explain.
c) Create a boxplot for these data.
d) Write a few sentences describing the distribution.

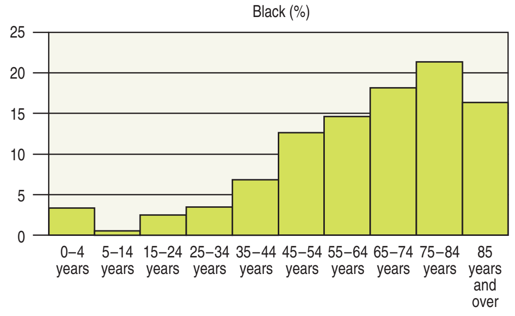
13. Hospital stays. The U.S. National Center for Health Statistics compiles data on the length of stay by patients in short-term hospitals and publishes its findings in *Vital and Health Statistics.* Data from a sample of 39 male patients and 35 female patients on length of stay (in days) are displayed in the histograms below.



a) What would you suggest be changed about these histograms to make them easier to compare?
b) Describe these distributions by writing a few sentences comparing the duration of hospitalization for men and women.
c) Can you suggest a reason for the peak in women's length of stay?

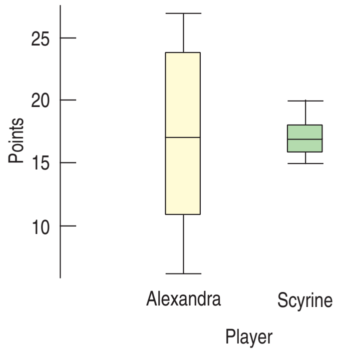
14. Deaths 2003. A National Vital Statistics Report (www.cdc.gov/nchs/) indicated that nearly 300,000 black Americans died in 2003, compared with just over 2 million white Americans. Here are histograms displaying the distributions of their ages at death:





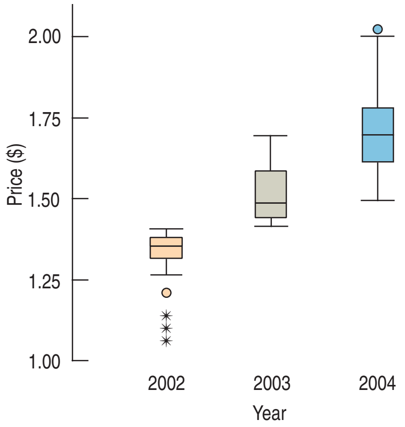
a) Describe the overall shapes of these distributions.
b) How do the distributions differ?
c) Look carefully at the bar definitions. Where do these plots violate the rules for statistical graphs?

15. Women's basketball. Here are boxplots of the points scored during the first 10 games of the season for both Scyrine and Alexandra:



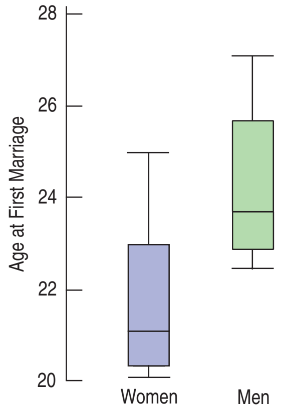
a) Summarize the similarities and differences in their performance so far.
b) The coach can take only one player to the state championship. Which one should she take? Why?

16. Gas prices. Here are boxplots of weekly gas prices at a service station in the midwestern United States (prices in $ per gallon):

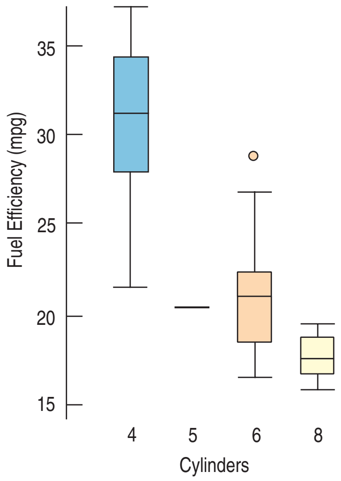


a) Compare the distribution of prices over the three years.
b) In which year were the prices least stable? Explain.

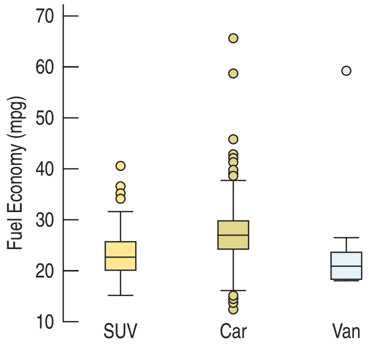
17. Marriage age. In 1975, did men and women marry at the same age? Here are boxplots of the age at first marriage for a sample of U.S. citizens then. Write a brief report discussing what these data show.



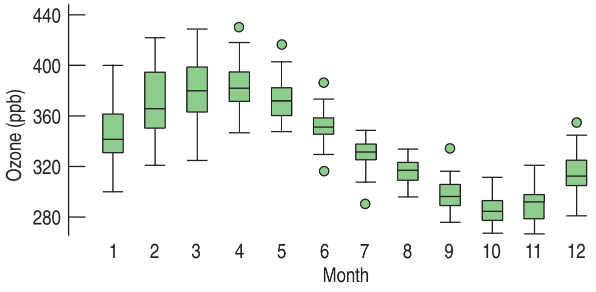
18. Fuel economy. Describe what these boxplots tell you about the relationship between the number of cylinders a car's engine has and the car's fuel economy (mpg):



19. Fuel economy II. The Environmental Protection Agency provides fuel economy and pollution information on over 2000 car models. Here is a boxplot of *Combined Fuel Economy* (using an average of driving conditions) in *miles per gallon* by vehicle *Type* (car, van, or SUV). Summarize what you see about the fuel economies of the three vehicle types.

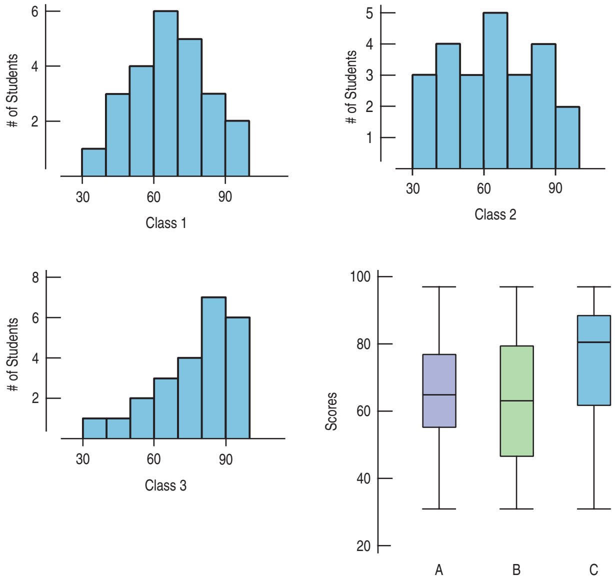


20. Ozone. Ozone levels (in parts per billion, ppb) were recorded at sites in New Jersey monthly between 1926 and 1971. Here are boxplots of the data for each month (over the 46 years), lined up in order (January = 1):



a) In what month was the highest ozone level ever recorded?
b) Which month has the largest IQR?
c) Which month has the smallest range?
d) Write a brief comparison of the ozone levels in January and June.
e) Write a report on the annual patterns you see in the ozone levels.

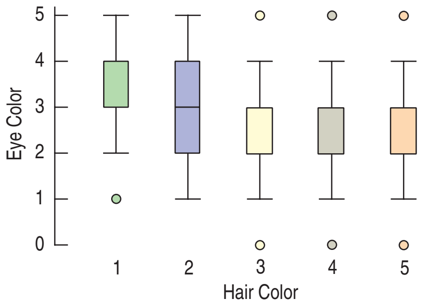
21. Test scores. Three Statistics classes all took the same test. Histograms and boxplots of the scores for each class are shown below. Match each class with the corresponding boxplot.



22. Eye and hair color. A survey of 1021 school-age children was conducted by randomly selecting children from several large urban elementary schools. Two of the questions concerned eye and hair color. In the survey, the following codes were used:

Hair Color	Eye Color
1 = Blond	1 = Blue
2 = Brown	2 = Green
3 = Black	3 = Brown
4 = Red	4 = Grey
5 = Other	5 = Other

The Statistics students analyzing the data were asked to study the relationship between eye and hair color. They produced this plot:



Is their graph appropriate? If so, summarize the findings. If not, explain why not.

23. Graduation? A survey of major universities asked what percentage of incoming freshmen usually graduate "on time" in 4 years. Use the summary statistics given to answer the questions that follow.

	% on Time
Count	48
Mean	68.35
Median	69.90
StdDev	10.20
Min	43.20
Max	87.40
Range	44.20
25th %tile	59.15
75th %tile	74.75

a) Would you describe this distribution as symmetric or skewed? Explain.
b) Are there any outliers? Explain.
c) Create a boxplot of these data.
d) Write a few sentences about the graduation rates.

24. Vineyards. Here are summary statistics for the sizes (in acres) of Finger Lakes vineyards:

Count	36
Mean	46.50 acres
StdDev	47.76
Median	33.50
IQR	36.50
Min	6
Q1	18.50
Q3	55
Max	250

a) Would you describe this distribution as symmetric or skewed? Explain.
b) Are there any outliers? Explain.
c) Create a boxplot of these data.
d) Write a few sentences about the sizes of the vineyards.

25. Caffeine. A student study of the effects of caffeine asked volunteers to take a memory test 2 hours after drinking soda. Some drank caffeine-free cola, some drank regular cola (with caffeine), and others drank a mixture of the two (getting a half-dose of caffeine). Here are the 5-number summaries for each group's scores (number of items recalled correctly) on the memory test:

	n	Min	Q1	Median	Q3	Max
No caffeine	15	16	20	21	24	26
Low caffeine	15	16	18	21	24	27
High caffeine	15	12	17	19	22	24

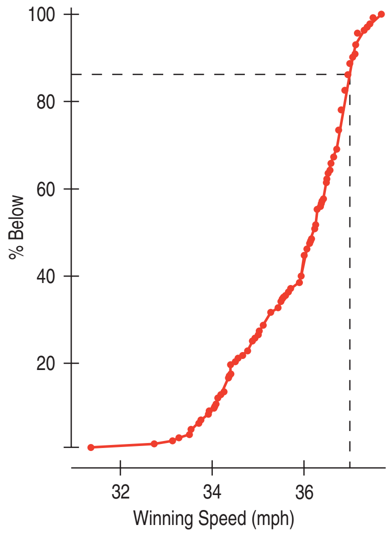
a) Describe the W's for these data.
b) Name the variables and classify each as categorical or quantitative.
c) Create parallel boxplots to display these results as best you can with this information.
d) Write a few sentences comparing the performances of the three groups.

26. SAT scores. Here are the summary statistics for Verbal SAT scores for a high school graduating class:

	n	Mean	Median	SD	Min	Max	Q1	Q3
Male	80	590	600	97.2	310	800	515	650
Female	82	602	625	102.0	360	770	530	680

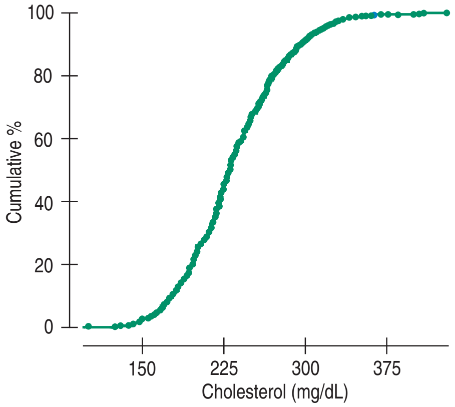
a) Create parallel boxplots comparing the scores of boys and girls as best you can from the information given.
b) Write a brief report on these results. Be sure to discuss the shape, center, and spread of the scores.

27. Derby speeds 2007. How fast do horses run? Kentucky Derby winners top 30 miles per hour, as shown in this graph. The graph shows the percentage of Derby winners that have run *slower* than each given speed. Note that few have won running less than 33 miles per hour, but about 86% of the winning horses have run less than 37 miles per hour. (A cumulative frequency graph like this is called an "ogive.")



a) Estimate the median winning speed.
b) Estimate the quartiles.
c) Estimate the range and the IQR.
d) Create a boxplot of these speeds.
e) Write a few sentences about the speeds of the Kentucky Derby winners.

28. **Cholesterol.** The Framingham Heart Study recorded the cholesterol levels of more than 1400 men. Here is an ogive of the distribution of these cholesterol measures. (An ogive shows the percentage of cases at or below a certain value.) Construct a boxplot for these data, and write a few sentences describing the distribution.



29. **Reading scores.** A class of fourth graders takes a diagnostic reading test, and the scores are reported by reading grade level. The 5-number summaries for the 14 boys and 11 girls are shown:

Boys:	2.0	3.9	4.3	4.9	6.0
Girls:	2.8	3.8	4.5	5.2	5.9

a) Which group had the highest score?
b) Which group had the greater range?
c) Which group had the greater interquartile range?
d) Which group's scores appear to be more skewed? Explain.
e) Which group generally did better on the test? Explain.
f) If the mean reading level for boys was 4.2 and for girls was 4.6, what is the overall mean for the class?

30. **Rainmakers?** In an experiment to determine whether seeding clouds with silver iodide increases rainfall, 52 clouds were randomly assigned to be seeded or not. The amount of rain they generated was then measured (in acre-feet). Here are the summary statistics:

	n	Mean	Median	SD	IQR	Q1	Q3
Unseeded	26	164.59	44.20	278.43	138.60	24.40	163
Seeded	26	441.98	221.60	650.79	337.60	92.40	430

a) Which of the summary statistics are most appropriate for describing these distributions. Why?
b) Do you see any evidence that seeding clouds may be effective? Explain.

31. **Industrial experiment.** Engineers at a computer production plant tested two methods for accuracy in drilling holes into a PC board. They tested how fast they could set the drilling machine by running 10 boards at each of two different speeds. To assess the results, they measured the distance (in inches) from the center of a target on the board to the center of the hole. The data and summary statistics are shown in the table:

	Distance (in.)	Speed		Distance (in.)	Speed
	0.000101	Fast		0.000098	Slow
	0.000102	Fast		0.000096	Slow
	0.000100	Fast		0.000097	Slow
	0.000102	Fast		0.000095	Slow
	0.000101	Fast		0.000094	Slow
	0.000103	Fast		0.000098	Slow
	0.000104	Fast		0.000096	Slow
	0.000102	Fast		0.975600	Slow
	0.000102	Fast		0.000097	Slow
	0.000100	Fast		0.000096	Slow
Mean	0.000102		Mean	0.097647	
StdDev	0.000001		StdDev	0.308481	

Write a report summarizing the findings of the experiment. Include appropriate visual and verbal displays of the distributions, and make a recommendation to the engineers if they are most interested in the accuracy of the method.

32. **Cholesterol.** A study examining the health risks of smoking measured the cholesterol levels of people who had smoked for at least 25 years and people of similar ages who had smoked for no more than 5 years and then stopped. Create appropriate graphical displays for both groups, and write a brief report comparing their cholesterol levels. Here are the data:

Smokers				Ex-Smokers		
225	211	209	284	250	134	300
258	216	196	288	249	213	310
250	200	209	280	175	174	328
225	256	243	200	160	188	321
213	246	225	237	213	257	292
232	267	232	216	200	271	227
216	243	200	155	238	163	263
216	271	230	309	192	242	249
183	280	217	305	242	267	243
287	217	246	351	217	267	218
200	280	209		217	183	228

33. **MPG.** A consumer organization compared gas mileage figures for several models of cars made in the United States with autos manufactured in other countries. The data are shown in the table:

Gas Mileage (mpg)	Country	Gas Mileage (mpg)	Country
16.9	U.S.	26.8	U.S.
15.5	U.S.	33.5	U.S.
19.2	U.S.	34.2	U.S.
18.5	U.S.	16.2	Other
30.0	U.S.	20.3	Other
30.9	U.S.	31.5	Other
20.6	U.S.	30.5	Other
20.8	U.S.	21.5	Other
18.6	U.S.	31.9	Other
18.1	U.S.	37.3	Other
17.0	U.S.	27.5	Other
17.6	U.S.	27.2	Other
16.5	U.S.	34.1	Other
18.2	U.S.	35.1	Other
26.5	U.S.	29.5	Other
21.9	U.S.	31.8	Other
27.4	U.S.	22.0	Other
28.4	U.S.	17.0	Other
28.8	U.S.	21.6	Other

a) Create graphical displays for these two groups.
b) Write a few sentences comparing the distributions.

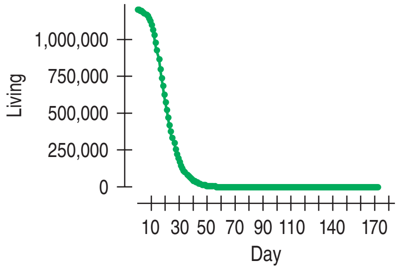
34. Baseball. American League baseball teams play their games with the designated hitter rule, meaning that pitchers do not bat. The League believes that replacing the pitcher, typically a weak hitter, with another player in the batting order produces more runs and generates more interest among fans. Following are the average number of runs scored in American League and National League stadiums for the first half of the 2001 season:

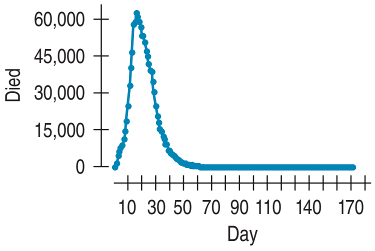
Average Runs	League	Average Runs	League
11.1	American	14.0	National
10.8	American	11.6	National
10.8	American	10.4	National
10.3	American	10.9	National
10.3	American	10.2	National
10.1	American	9.5	National
10.0	American	9.5	National
9.5	American	9.5	National
9.4	American	9.5	National
9.3	American	9.1	National
9.2	American	8.8	National
9.2	American	8.4	National
9.0	American	8.3	National
8.3	American	8.2	National
		8.1	National
		7.9	National

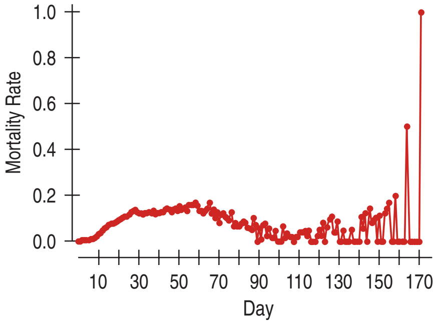
a) Create an appropriate graphical display of these data.
b) Write a few sentences comparing the average number of runs scored per game in the two leagues. (Remember: shape, center, spread, unusual features!)

c) Coors Field in Denver stands a mile above sea level, an altitude far greater than that of any other major league ball park. Some believe that the thinner air makes it harder for pitchers to throw curveballs and easier for batters to hit the ball a long way. Do you see any evidence that the 14 runs scored per game there is unusually high? Explain.

35. Fruit Flies. Researchers tracked a population of 1,203,646 fruit flies, counting how many died each day for 171 days. Here are three timeplots offering different views of these data. One shows the number of flies alive on each day, one the number who died that day, and the third the mortality rate—the fraction of the number alive who died. On the last day studied, the last 2 flies died, for a mortality rate of 1.0.







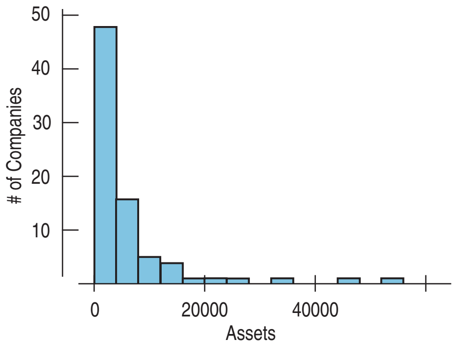
a) On approximately what day did the most flies die?
b) On what day during the first 100 days did the largest *proportion* of flies die?
c) When did the number of fruit flies alive stop changing very much from day to day?

36. Drunk driving 2005. Accidents involving drunk drivers account for about 40% of all deaths on the nation's highways. The table tracks the number of alcohol-related fatalities for 24 years. (www.madd.org)

Year	Deaths (thousands)	Year	Deaths (thousands)
1982	26.2	1994	17.3
1983	24.6	1995	17.7
1984	24.8	1996	17.7
1985	23.2	1997	16.7
1986	25.0	1998	16.7
1987	24.1	1999	16.6
1988	23.8	2000	17.4
1989	22.4	2001	17.4
1990	22.6	2002	17.5
1991	20.2	2003	17.1
1992	18.3	2004	16.9
1993	17.9	2005	16.9

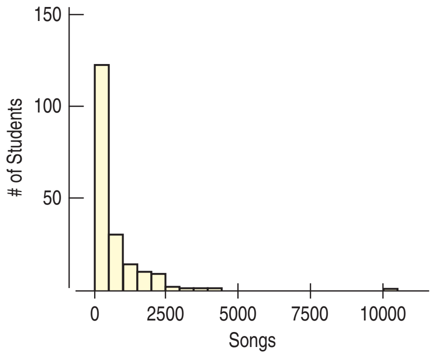
a) Create a stem-and-leaf display or a histogram of these data.
b) Create a timeplot.
c) Using features apparent in the stem-and-leaf display (or histogram) and the timeplot, write a few sentences about deaths caused by drunk driving.

T **37. Assets.** Here is a histogram of the assets (in millions of dollars) of 79 companies chosen from the *Forbes* list of the nation's top corporations:



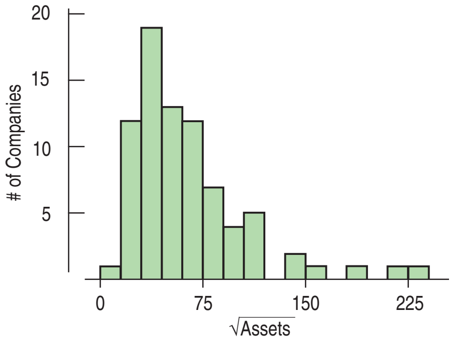
a) What aspect of this distribution makes it difficult to summarize, or to discuss, center and spread?
b) What would you suggest doing with these data if we want to understand them better?

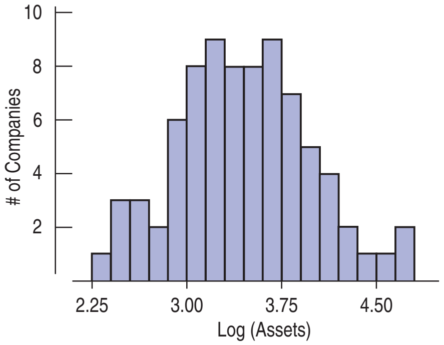
38. Music library. Students were asked how many songs they had in their digital music libraries. Here's a display of the responses:



a) What aspect of this distribution makes it difficult to summarize, or to discuss, center and spread?
b) What would you suggest doing with these data if we want to understand them better?

T **39. Assets again.** Here are the same data you saw in Exercise 37 after re-expressions as the square root of assets and the logarithm of assets:





a) Which re-expression do you prefer? Why?
b) In the square root re-expression, what does the value 50 actually indicate about the company's assets?
c) In the logarithm re-expression, what does the value 3 actually indicate about the company's assets?

T **40. Rainmakers.** The table lists the amount of rainfall (in acre-feet) from the 26 clouds seeded with silver iodide discussed in Exercise 30:

2745	703	302	242	119	40	7
1697	489	274	200	118	32	4
1656	430	274	198	115	31	
978	334	255	129	92	17	

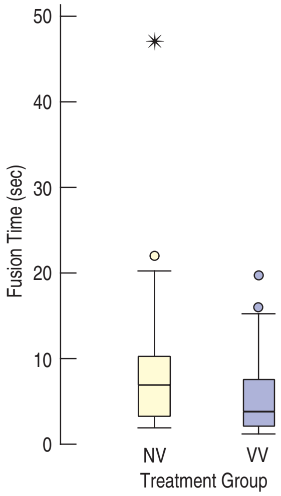
a) Why is acre-feet a good way to measure the amount of precipitation produced by cloud seeding?
b) Plot these data, and describe the distribution.
c) Create a re-expression of these data that produces a more advantageous distribution.
d) Explain what your re-expressed scale means.

T **41. Stereograms.** Stereograms appear to be composed entirely of random dots. However, they contain separate images that a viewer can "fuse" into a three-dimensional (3D) image by staring at the dots while defocusing the eyes. An experiment was performed to determine whether knowledge of the embedded image affected the time required for subjects to fuse the images. One group of subjects (group NV) received no information or just verbal information about the shape of the embedded object. A second group (group VV) received both verbal information and visual information (specifically, a drawing of the object). The experimenters measured how many seconds it took for the subject to report that he or she saw the 3D image.

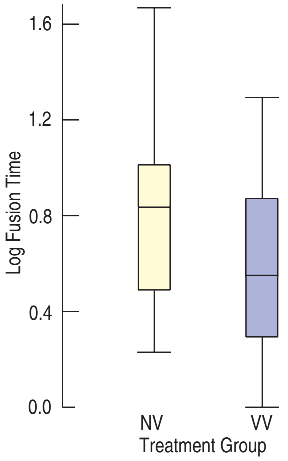
a) What two variables are discussed in this description?

b) For each variable, is it quantitative or categorical? If quantitative, what are the units?

c) The boxplots compare the fusion times for the two treatment groups. Write a few sentences comparing these distributions. What does the experiment show?



T **42. Stereograms, revisited.** Because of the skewness of the distributions of fusion times described in Exercise 41, we might consider a re-expression. Here are the boxplots of the *log* of fusion times. Is it better to analyze the original fusion times or the log fusion times? Explain.



JUST CHECKING Answers

1. The % late arrivals have a unimodal, symmetric distribution centered at about 20%. In most months between 16% and 23% of the flights arrived late.
2. The boxplot of % late arrivals makes it easier to see that the median is just below 20%, with quartiles at about 17% and 22%. It nominates two months as high outliers.
3. The boxplots by month show a strong seasonal pattern. Flights are more likely to be late in the winter and summer and less likely to be late in the spring and fall. One likely reason for the pattern is snowstorms in the winter and thunderstorms in the summer.

CHAPTER

6

The Standard Deviation as a Ruler and the Normal Model

The women's heptathlon in the Olympics consists of seven track and field events: the 200-m and 800-m runs, 100-m high hurdles, shot put, javelin, high jump, and long jump. To determine who should get the gold medal, somehow the performances in all seven events have to be combined into one score. How can performances in such different events be compared? They don't even have the same units; the races are recorded in minutes and seconds and the throwing and jumping events in meters. In the 2004 Olympics, Austra Skujyté of Lithuania put the shot 16.4 meters, about 3 meters farther than the average of all contestants. Carolina Klüft won the long jump with a 6.78-m jump, about a meter better than the average. Which performance deserves more points? Even though both events are measured in meters, it's not clear how to compare them. The solution to the problem of how to compare scores turns out to be a useful method for comparing all sorts of values whether they have the same units or not.

The Standard Deviation as a Ruler

Grading on a Curve
If you score 79% on an exam, what grade should you get? One teaching philosophy looks only at the raw percentage, 79, and bases the grade on that alone. Another looks at your *relative* performance and bases the grade on how you did compared with the rest of the class. Teachers and students still debate which method is better.

The trick in comparing very different-looking values is to use standard deviations. The standard deviation tells us how the whole collection of values varies, so it's a natural ruler for comparing an individual value to the group. Over and over during this course, we will ask questions such as "How far is this value from the mean?" or "How different are these two statistics?" The answer in every case will be to measure the distance or difference in standard deviations.

The concept of the standard deviation as a ruler is not special to this course. You'll find statistical distances measured in standard deviations throughout Statistics, up to the most advanced levels.[1] This approach is one of the basic tools of statistical thinking.

[1] Other measures of spread could be used as well, but the standard deviation is the most common measure, and it is almost always used as the ruler.

In order to compare the two events, let's start with a picture. This time we'll use stem-and-leaf displays so we can see the individual distances.

Long Jump Stem	Leaf	Shot Put Stem	Leaf
67	8	16	4
66		15	
65	1	15	
64	2	14	56778
63	0566	14	24
62	11235	13	5789
61	0569	13	012234
60	2223	12	55
59	0278	12	0144
58	4	11	59
57	0	11	23

FIGURE 6.1

Stem-and-leaf displays for both the long jump and the shot put in the 2004 Olympic Heptathlon. Carolina Klüft (green scores) won the long jump, and Austra Skujyté (red scores) won the shot put. Which heptathlete did better for both events combined?

The two winning performances on the top of each stem-and-leaf display appear to be about the same distance from the center of the pack. But look again carefully. What do we mean by the *same distance*? The two displays have different scales. Each line in the stem-and-leaf for the shot put represents half a meter, but for the long jump each line is only a tenth of a meter. It's only because our eyes naturally adjust the scales and use the standard deviation as the ruler that we see each as being about the same distance from the center of the data. How can we make this hunch more precise? Let's see how many standard deviations each performance is from the mean.

Klüft's 6.78-m long jump is 0.62 meters longer than the mean jump of 6.16 m. How many *standard deviations* better than the mean is that? The standard deviation for this event was 0.23 m, so her jump was (6.78 − 6.16)/0.23 = 0.62/0.23 = 2.70 *standard deviations better* than the mean. Skujyté's winning shot put was 16.40 − 13.29 = 3.11 meters longer than the mean shot put distance, and that's 3.11/1.24 = 2.51 standard deviations better than the mean. That's a great performance but not quite as impressive as Klüft's long jump, which was farther above the mean, as measured in *standard deviations.*

	Event	
	Long Jump	Shot Put
Mean (all contestants)	6.16 m	13.29 m
SD	0.23 m	1.24 m
n	26	28
Klüft	6.78 m	14.77 m
Skujyté	6.30 m	16.40 m

Standardizing with *z*-Scores

NOTATION ALERT:

There goes another letter. We always use the letter z to denote values that have been standardized with the mean and standard deviation.

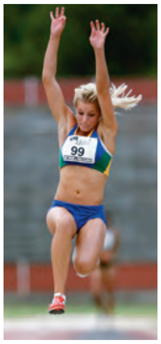
To compare these athletes' performances, we determined how many standard deviations from the event's mean each was.

Expressing the distance in standard deviations ***standardizes*** the performances. To standardize a value, we simply subtract the mean performance in that event and then divide this difference by the standard deviation. We can write the calculation as

$$z = \frac{y - \bar{y}}{s}.$$

These values are called **standardized values,** and are commonly denoted with the letter z. Usually, we just call them **z-scores.**

Standardized values have *no units*. z-scores measure the distance of each data value from the mean in standard deviations. A z-score of 2 tells us that a data value is 2 standard deviations above the mean. It doesn't matter whether the original variable was measured in inches, dollars, or seconds. Data values below the mean have negative z-scores, so a z-score of −1.6 means that the data value was 1.6 standard deviations below the mean. Of course, regardless of the direction, the farther a data value is from the mean, the more unusual it is, so a z-score of −1.3



is more extraordinary than a *z*-score of 1.2. Looking at the *z*-scores, we can see that even though both were winning scores, Klüft's long jump with a *z*-score of 2.70 is slightly more impressive than Skujyté's shot put with a *z*-score of 2.51.

FOR EXAMPLE Standardizing skiing times

The men's combined skiing event in the winter Olympics consists of two races: a downhill and a slalom. Times for the two events are added together, and the skier with the lowest total time wins. In the 2006 Winter Olympics, the mean slalom time was 94.2714 seconds with a standard deviation of 5.2844 seconds. The mean downhill time was 101.807 seconds with a standard deviation of 1.8356 seconds. Ted Ligety of the United States, who won the gold medal with a combined time of 189.35 seconds, skied the slalom in 87.93 seconds and the downhill in 101.42 seconds.

Question: On which race did he do better compared with the competition?

For the slalom, Ligety's *z*-score is found by subtracting the mean time from his time and then dividing by the standard deviation:

$$z_{Slalom} = \frac{87.93 - 94.2714}{5.2844} = -1.2$$

Similarly, his *z*-score for the downhill is:

$$z_{Downhill} = \frac{101.42 - 101.807}{1.8356} = -0.21$$

The *z*-scores show that Ligety's time in the slalom is farther below the mean than his time in the downhill. His performance in the slalom was more remarkable.

By using the standard deviation as a ruler to measure statistical distance from the mean, we can compare values that are measured on different variables, with different scales, with different units, or for different individuals. To determine the winner of the heptathlon, the judges must combine performances on seven very different events. Because they want the score to be absolute, and *not* dependent on the particular athletes in each Olympics, they use predetermined tables, but they could combine scores by standardizing each, and then adding the *z*-scores together to reach a total score. The only trick is that they'd have to switch the sign of the *z*-score for running events, because unlike throwing and jumping, it's better to have a running time below the mean (with a negative *z*-score).

To combine the scores Skujyté and Klüft earned in the long jump and the shot put, we standardize both events as shown in the table. That gives Klüft her 2.70 *z*-score in the long jump and a 1.19 in the shot put, for a total of 3.89. Skujyté's shot put gave her a 2.51, but her long jump was only 0.61 SDs above the mean, so her total is 3.12.

Is this the result we wanted? Yes. Each won one event, but Klüft's shot put was second best, while Skujyté's long jump was seventh. The *z*-scores measure how far each result is from the event mean in standard deviation units. And because they are both in standard deviation units, we can combine them. Not coincidentally, Klüft went on to win the gold medal for the entire seven-event heptathlon, while Skujyté got the silver.

		Event	
		Long Jump	Shot Put
	Mean	6.16 m	13.29 m
	SD	0.23 m	1.24 m
Klüft	Performance	6.78 m	14.77 m
	z-score	$\frac{6.78 - 6.16}{0.23} = 2.70$	$\frac{14.77 - 13.29}{1.24} = 1.19$
	Total *z*-score	2.70 + 1.19 = **3.89**	
Skujyté	Performance	6.30 m	16.40 m
	z-score	$\frac{6.30 - 6.16}{0.23} = 0.61$	$\frac{16.40 - 13.29}{1.24} = 2.51$
	Total *z*-score	0.61 + 2.51 = **3.12**	

FOR EXAMPLE **Combining z-scores**

In the 2006 winter Olympics men's combined event, Ivica Kostelić of Croatia skied the slalom in 89.44 seconds and the downhill in 100.44 seconds. He thus beat Ted Ligety in the downhill, but not in the slalom. Maybe he should have won the gold medal.

Question: Considered in terms of standardized scores, which skier did better?

Kostelić's z-scores are:

$$z_{Slalom} = \frac{89.44 - 94.2714}{5.2844} = -0.91 \quad \text{and} \quad z_{Downhill} = \frac{100.44 - 101.807}{1.8356} = -0.74$$

The sum of his z-scores is approximately –1.65. Ligety's z-score sum is only about –1.41. Because the standard deviation of the downhill times is so much smaller, Kostelić's better performance there means that he would have won the event if standardized scores were used.

When we standardize data to get a *z*-score, we do two things. First, we shift the data by subtracting the mean. Then, we rescale the values by dividing by their standard deviation. We often shift and rescale data. What happens to a grade distribution if *everyone* gets a five-point bonus? Everyone's grade goes up, but does the shape change? (*Hint:* Has anyone's distance from the mean changed?) If we switch from feet to meters, what happens to the distribution of heights of students in your class? Even though your intuition probably tells you the answers to these questions, we need to look at exactly how shifting and rescaling work.

JUST CHECKING

1. Your Statistics teacher has announced that the lower of your two tests will be dropped. You got a 90 on test 1 and an 80 on test 2. You're all set to drop the 80 until she announces that she grades "on a curve." She standardized the scores in order to decide which is the lower one. If the mean on the first test was 88 with a standard deviation of 4 and the mean on the second was 75 with a standard deviation of 5,

a) Which one will be dropped?
b) Does this seem "fair"?

Shifting Data

Since the 1960s, the Centers for Disease Control's National Center for Health Statistics has been collecting health and nutritional information on people of all ages and backgrounds. A recent survey, the National Health and Nutrition Examination Survey (NHANES) 2001–2002,[2] measured a wide variety of variables, including body measurements, cardiovascular fitness, blood chemistry, and demographic information on more than 11,000 individuals.

[2] www.cdc.gov/nchs/nhanes.htm

WHO	80 male participants of the NHANES survey between the ages of 19 and 24 who measured between 68 and 70 inches tall
WHAT	Their weights
UNIT	Kilograms
WHEN	2001–2002
WHERE	United States
WHY	To study nutrition, and health issues and trends
HOW	National survey

Included in this group were 80 men between 19 and 24 years old of average height (between 5′8″ and 5′10″ tall). Here are a histogram and boxplot of their weights:

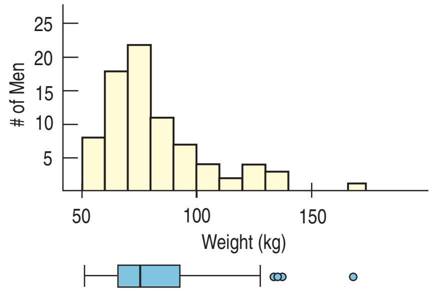


FIGURE 6.2

Histogram and boxplot for the men's weights. The shape is skewed to the right with several high outliers.

Their mean weight is 82.36 kg. For this age and height group, the National Institutes of Health recommends a maximum healthy weight of 74 kg, but we can see that some of the men are heavier than the recommended weight. To compare their weights to the recommended maximum, we could subtract 74 kg from each of their weights. What would that do to the center, shape, and spread of the histogram? Here's the picture:

A S ***Activity:*** **Changing the Baseline.** What happens when we shift data? Do measures of center and spread change?

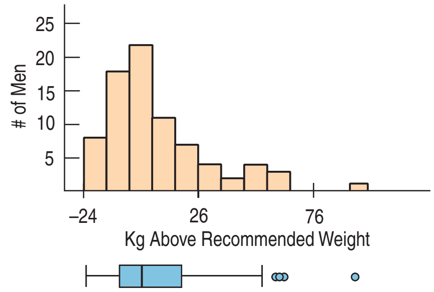


FIGURE 6.3

Subtracting 74 kilograms shifts the entire histogram down but leaves the spread and the shape exactly the same.

Doctors' height and weight charts sometimes give ideal weights for various heights that include 2-inch heels. If the mean height of adult women is 66 inches including 2-inch heels, what is the mean height of women without shoes? Each woman is shorter by 2 inches when barefoot, so the mean is decreased by 2 inches, to 64 inches.

On average, they weigh 82.36 kg, so on average they're 8.36 kg overweight. And, after subtracting 74 from each weight, the mean of the new distribution is $82.36 - 74 = 8.36$ kg. In fact, when we **shift** the data by adding (or subtracting) a constant to each value, all measures of position (center, percentiles, min, max) will increase (or decrease) by the same constant.

What about the spread? What does adding or subtracting a constant value do to the spread of the distribution? Look at the two histograms again. Adding or subtracting a constant changes each data value equally, so the entire distribution just shifts. Its shape doesn't change and neither does the spread. None of the measures of spread we've discussed—not the range, not the IQR, not the standard deviation—changes.

> *Adding (or subtracting) a constant to every data value adds (or subtracts) the same constant to measures of position, but leaves measures of spread unchanged.*

Rescaling Data

Not everyone thinks naturally in metric units. Suppose we want to look at the weights in pounds instead. We'd have to **rescale** the data. Because there are about 2.2 pounds in every kilogram, we'd convert the weights by multiplying each value by 2.2. Multiplying or dividing each value by a constant changes the measurement units. Here are histograms of the two weight distributions, plotted on the same scale, so you can see the effect of multiplying:

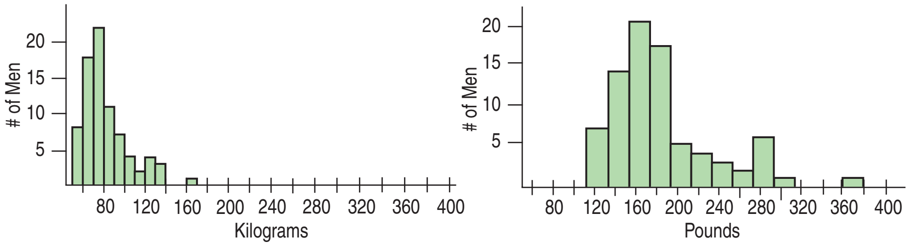


FIGURE 6.4

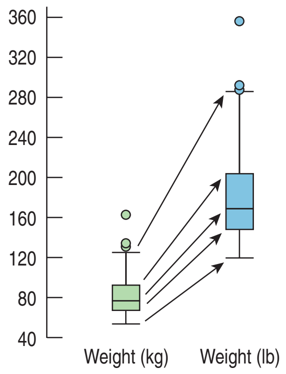
Men's weights in both kilograms and pounds. How do the distributions and numerical summaries change?

***Simulation:* Changing the Units.** Change the center and spread values for a distribution and watch the summaries change (or not, as the case may be).

What happens to the shape of the distribution? Although the histograms don't look exactly alike, we see that the shape really hasn't changed: Both are unimodal and skewed to the right.

What happens to the mean? Not too surprisingly, it gets multiplied by 2.2 as well. The men weigh 82.36 kg on average, which is 181.19 pounds. As the boxplots and 5-number summaries show, all measures of position act the same way. They all get multiplied by this same constant.

What happens to the spread? Take a look at the boxplots. The spread in pounds (on the right) is larger. How much larger? If you guessed 2.2 times, you've figured out how measures of spread get rescaled.



	Weight (kg)	Weight (lb)
Min	54.3	119.46
Q1	67.3	148.06
Median	76.85	169.07
Q3	92.3	203.06
Max	161.5	355.30
IQR	25	55
SD	22.27	48.99

FIGURE 6.5

The boxplots (drawn on the same scale) show the weights measured in kilograms (on the left) and pounds (on the right). Because 1 kg is 2.2 lb, all the points in the right box are 2.2 times larger than the corresponding points in the left box. So each measure of position and spread is 2.2 times as large when measured in pounds rather than kilograms.

When we multiply (or divide) all the data values by any constant, all measures of position (such as the mean, median, and percentiles) and measures of spread (such as the range, the IQR, and the standard deviation) are multiplied (or divided) by that same constant.

FOR EXAMPLE Rescaling the slalom

Recap: The times in the men's combined event at the winter Olympics are reported in minutes and seconds. Previously, we converted these to seconds and found the mean and standard deviation of the slalom times to be 94.2714 seconds and 5.2844 seconds, respectively.

Question: Suppose instead that we had reported the times in minutes—that is, that each individual time was divided by 60. What would the resulting mean and standard deviation be?

Dividing all the times by 60 would divide both the mean and the standard deviation by 60:

$$\text{Mean} = 94.2714/60 = 1.5712 \text{ minutes}; \quad \text{SD} = 5.2844/60 = 0.0881 \text{ minutes}.$$

JUST CHECKING

2. In 1995 the Educational Testing Service (ETS) adjusted the scores of SAT tests. Before ETS recentered the SAT Verbal test, the mean of all test scores was 450.

a) How would adding 50 points to each score affect the mean?

b) The standard deviation was 100 points. What would the standard deviation be after adding 50 points?

c) Suppose we drew boxplots of test takers' scores a year before and a year after the recentering. How would the boxplots of the two years differ?

3. A company manufactures wheels for in-line skates. The diameters of the wheels have a mean of 3 inches and a standard deviation of 0.1 inches. Because so many of their customers use the metric system, the company decided to report their production statistics in millimeters (1 inch = 25.4 mm). They report that the standard deviation is now 2.54 mm. A corporate executive is worried about this increase in variation. Should he be concerned? Explain.

Back to *z*-scores

A S ***Activity:*** **Standardizing.** What if we both shift and rescale? The result is so nice that we give it a name.

Standardizing data into *z*-scores is just shifting them by the mean and rescaling them by the standard deviation. Now we can see how standardizing affects the distribution. When we subtract the mean of the data from every data value, we shift the mean to zero. As we have seen, such a shift doesn't change the standard deviation.

When we *divide* each of these shifted values by *s*, however, the standard deviation should be divided by *s* as well. Since the standard deviation was *s* to start with, the new standard deviation becomes 1.

How, then, does standardizing affect the distribution of a variable? Let's consider the three aspects of a distribution: the shape, center, and spread.

_z_-scores have mean 0 and standard deviation 1.

- *Standardizing into z-scores does not change the* ***shape*** *of the distribution of a variable.*
- *Standardizing into z-scores changes the* ***center*** *by making the mean 0.*
- *Standardizing into z-scores changes the* ***spread*** *by making the standard deviation 1.*

STEP-BY-STEP EXAMPLE Working with Standardized Variables

Many colleges and universities require applicants to submit scores on standardized tests such as the SAT Writing, Math, and Critical Reading (Verbal) tests. The college your little sister wants to apply to says that while there is no minimum score required, the middle 50% of their students have combined SAT scores between 1530 and 1850. You'd feel confident if you knew her score was in their top 25%, but unfortunately she took the ACT test, an alternative standardized test.

Question: How high does her ACT need to be to make it into the top quarter of equivalent SAT scores?

To answer that question you'll have to standardize all the scores, so you'll need to know the mean and standard deviations of scores for some group on both tests. The college doesn't report the mean or standard deviation for their applicants on either test, so we'll use the group of all test takers nationally. For college-bound seniors, the average combined SAT score is about 1500 and the standard deviation is about 250 points. For the same group, the ACT average is 20.8 with a standard deviation of 4.8.

THINK	**Plan** State what you want to find out. **Variables** Identify the variables and report the W's (if known). Check the appropriate conditions.	I want to know what ACT score corresponds to the upper-quartile SAT score. I know the mean and standard deviation for both the SAT and ACT scores based on all test takers, but I have no individual data values. ✓ **Quantitative Data Condition:** Scores for both tests are quantitative but have no meaningful units other than points.
SHOW	**Mechanics** Standardize the variables. The y-value we seek is z standard deviations above the mean.	The middle 50% of SAT scores at this college fall between 1530 and 1850 points. To be in the top quarter, my sister would have to have a score of at least 1850. That's a z-score of $z = \frac{(1850 - 1500)}{250} = 1.40$ So an SAT score of 1850 is 1.40 standard deviations above the mean of all test takers. For the ACT, 1.40 standard deviations above the mean is $20.8 + 1.40(4.8) = 27.52$.
TELL	**Conclusion** Interpret your results in context.	To be in the top quarter of applicants in terms of combined SAT score, she'd need to have an ACT score of at least 27.52.

When Is a *z*-score BIG?

A z-score gives us an indication of how unusual a value is because it tells us how far it is from the mean. If the data value sits right at the mean, it's not very far at all and its z-score is 0. A z-score of 1 tells us that the data value is 1 standard deviation above the mean, while a z-score of -1 tells us that the value is 1 standard deviation below the mean. How far from 0 does a z-score have to be to be interesting or unusual? There is no universal standard, but the larger the score is (negative or positive), the more unusual it is. We know that 50% of the data lie between the quartiles. For symmetric data, the standard deviation is usually a bit smaller than the IQR, and it's not uncommon for at least half of the data to have z-scores between -1 and 1. But no matter what the shape of the distribution, a z-score of 3 (plus or minus) or more is rare, and a z-score of 6 or 7 shouts out for attention.

Is Normal Normal?

Don't be misled. The name "Normal" doesn't mean that these are the *usual* shapes for histograms. The name follows a tradition of positive thinking in Mathematics and Statistics in which functions, equations, and relationships that are easy to work with or have other nice properties are called "normal", "common", "regular", "natural", or similar terms. It's as if by calling them ordinary, we could make them actually occur more often and simplify our lives.

To say more about how big we expect a z-score to be, we need to *model* the data's distribution. A model will let us say much more precisely how often we'd be likely to see z-scores of different sizes. Of course, like all models of the real world, the model will be wrong—wrong in the sense that it can't match

"All models are wrong—but some are useful."

—George Box, famous statistician

reality exactly. But it can still be useful. Like a physical model, it's something we can look at and manipulate in order to learn more about the real world.

Models help our understanding in many ways. Just as a model of an airplane in a wind tunnel can give insights even though it doesn't show every rivet,[3] models of data give us summaries that we can learn from and use, even though they don't fit each data value exactly. It's important to remember that they're only *models* of reality and not reality itself. But without models, what we can learn about the world at large is limited to only what we can say about the data we have at hand.

There is no universal standard for z-scores, but there is a model that shows up over and over in Statistics. You may have heard of "bell-shaped curves." Statisticians call them Normal models. **Normal models** are appropriate for distributions whose shapes are unimodal and roughly symmetric. For these distributions, they provide a measure of how extreme a z-score is. Fortunately, there is a Normal model for every possible combination of mean and standard deviation. We write $N(\mu, \sigma)$ to represent a Normal model with a mean of μ and a standard deviation of σ. Why the Greek? Well, *this* mean and standard deviation are not numerical summaries of data. They are part of the model. They don't come from the data. Rather, they are numbers that we choose to help specify the model. Such numbers are called **parameters** of the model.

NOTATION ALERT:

$N(\mu, \sigma)$ always denotes a Normal model. The μ, pronounced "mew," is the Greek letter for "m" and always represents the mean in a model. The σ, sigma, is the lowercase Greek letter for "s" and always represents the standard deviation in a model.

We don't want to confuse the parameters with summaries of the data such as $\bar{y}$ and s, so we use special symbols. In Statistics, we almost always use Greek letters for parameters. By contrast, summaries of data are called **statistics** and are usually written with Latin letters.

If we model data with a Normal model and standardize them using the corresponding μ and σ, we still call the standardized value a **z-score,** and we write

$$z = \frac{y - \mu}{\sigma}.$$

Usually it's easier to standardize data first (using its mean and standard deviation). Then we need only the model $N(0,1)$. The Normal model with mean 0 and standard deviation 1 is called the **standard Normal model** (or the **standard Normal distribution**).

Is the Standard Normal a standard?

Yes. We call it the "Standard Normal" because it models standardized values. It is also a "standard" because this is the particular Normal model that we almost always use.

But be careful. You shouldn't use a Normal model for just any data set. Remember that standardizing won't change the shape of the distribution. If the distribution is not unimodal and symmetric to begin with, standardizing won't make it Normal.

When we use the Normal model, we assume that the distribution of the data is, well, Normal. Practically speaking, there's no way to check whether this **Normality Assumption** is true. In fact, it almost certainly is not true. Real data don't behave like mathematical models. Models are idealized; real data are real. The good news, however, is that to use a Normal model, it's sufficient to check the following condition:

A S ***Activity:*** **Working with Normal Models.** Learn more about the Normal model and see what data drawn at random from a Normal model might look like.

> **Nearly Normal Condition.** The shape of the data's distribution is unimodal and symmetric. Check this by making a histogram (or a Normal probability plot, which we'll explain later).

Don't model data with a Normal model without checking whether the condition is satisfied.

All models make **assumptions.** Whenever we model—and we'll do that often—we'll be careful to point out the assumptions that we're making. And, what's even more important, we'll check the associated **conditions** in the data to make sure that those assumptions are reasonable.

[3] In fact, the model is useful *because* it doesn't have every rivet. It is because models offer a simpler view of reality that they are so useful as we try to understand reality.

The 68–95–99.7 Rule

One in a Million

These magic 68, 95, 99.7 values come from the Normal model. As a model, it can give us corresponding values for any *z*-score. For example, it tells us that fewer than 1 out of a million values have *z*-scores smaller than −5.0 or larger than +5.0. So if someone tells you you're "one in a million," they must really admire your *z*-score.

TI-*nspire*

The 68–95–99.7 Rule. See it work for yourself.

Normal models give us an idea of how extreme a value is by telling us how likely it is to find one that far from the mean. We'll soon show how to find these numbers precisely—but one simple rule is usually all we need.

It turns out that in a Normal model, about 68% of the values fall within 1 standard deviation of the mean, about 95% of the values fall within 2 standard deviations of the mean, and about 99.7%—almost all—of the values fall within 3 standard deviations of the mean. These facts are summarized in a rule that we call (let's see . . .) the **68–95–99.7 Rule.**[4]

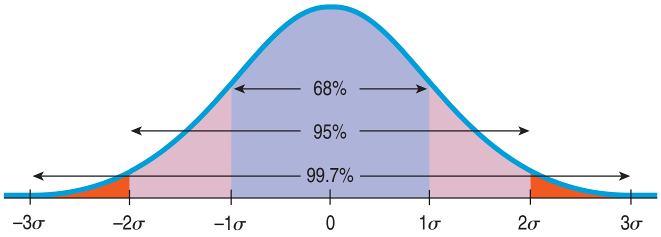


Figure 6.6

Reaching out one, two, and three standard deviations on a Normal model gives the 68−95−99.7 Rule, seen as proportions of the area under the curve.

FOR EXAMPLE Using the 68–95–99.7 Rule

Question: In the 2006 Winter Olympics men's combined event, Jean-Baptiste Grange of France skied the slalom in 88.46 seconds—about 1 standard deviation faster than the mean. If a Normal model is useful in describing slalom times, about how many of the 35 skiers finishing the event would you expect skied the slalom *faster* than Jean-Baptiste?

From the 68–95–99.7 Rule, we expect 68% of the skiers to be within one standard deviation of the mean. Of the remaining 32%, we expect half on the high end and half on the low end. 16% of 35 is 5.6, so, conservatively, we'd expect about 5 skiers to do better than Jean-Baptiste.

JUST CHECKING

4. As a group, the Dutch are among the tallest people in the world. The average Dutch man is 184 cm tall—just over 6 feet (and the average Dutch woman is 170.8 cm tall—just over 5′7″). If a Normal model is appropriate and the standard deviation for men is about 8 cm, what percentage of all Dutch men will be over 2 meters (6′6″) tall?

5. Suppose it takes you 20 minutes, on average, to drive to school, with a standard deviation of 2 minutes. Suppose a Normal model is appropriate for the distributions of driving times.

a) How often will you arrive at school in less than 22 minutes?

b) How often will it take you more than 24 minutes?

c) Do you think the distribution of your driving times is unimodal and symmetric?

d) What does this say about the accuracy of your predictions? Explain.

[4] This rule is also called the "Empirical Rule" because it originally came from observation. The rule was first published by Abraham de Moivre in 1733, 75 years before the Normal model was discovered. Maybe it should be called "de Moivre's Rule," but that wouldn't help us remember the important numbers, 68, 95, and 99.7.

The First Three Rules for Working with Normal Models

A S ***Activity:* Working with Normal Models.** Well, actually playing with them. This interactive tool lets you do what this chapter's figures can't do, move them!

1. Make a picture.
2. Make a picture.
3. Make a picture.

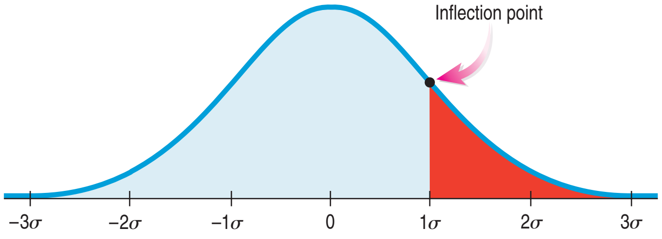
Although we're thinking about models, not histograms of data, the three rules don't change. To help you think clearly, a simple hand-drawn sketch is all you need. Even experienced statisticians sketch pictures to help them think about Normal models. You should too.

Of course, when we have data, we'll also need to make a histogram to check the **Nearly Normal Condition** to be sure we can use the Normal model to model the data's distribution. Other times, we may be told that a Normal model is appropriate based on prior knowledge of the situation or on theoretical considerations.

A S ***Activity:* Normal Models.** Normal models have several interesting properties—see them here.

How to Sketch a Normal Curve That Looks Normal To sketch a good Normal curve, you need to remember only three things:

- The Normal curve is bell-shaped and symmetric around its mean. Start at the middle, and sketch to the right and left from there.
- Even though the Normal model extends forever on either side, you need to draw it only for 3 standard deviations. After that, there's so little left that it isn't worth sketching.
- The place where the bell shape changes from curving downward to curving back up—the *inflection point*—is exactly one standard deviation away from the mean.



STEP-BY-STEP EXAMPLE Working with the 68–95–99.7 Rule

The SAT Reasoning Test has three parts: Writing, Math, and Critical Reading (Verbal). Each part has a distribution that is roughly unimodal and symmetric and is designed to have an overall mean of about 500 and a standard deviation of 100 for all test takers. In any one year, the mean and standard deviation may differ from these target values by a small amount, but they are a good overall approximation.

Question: Suppose you earned a 600 on one part of your SAT. Where do you stand among all students who took that test?

You could calculate your z-score and find out that it's $z = (600 - 500)/100 = 1.0$, but what does that tell you about your percentile? You'll need the Normal model and the 68–95–99.7 Rule to answer that question.

THINK

Plan State what you want to know.

I want to see how my SAT score compares with the scores of all other students. To do that, I'll need to model the distribution.

Variables Identify the variable and report the W's.

Let y = my SAT score. Scores are quantitative but have no meaningful units other than points.

Be sure to check the appropriate conditions.

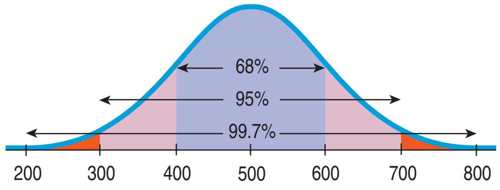
✓ **Nearly Normal Condition:** If I had data, I would check the histogram. I have no data, but I am told that the SAT scores are roughly unimodal and symmetric.

Specify the parameters of your model.

I will model SAT score with a *N*(500, 100) model.

SHOW

Mechanics Make a picture of this Normal model. (A simple sketch is all you need.)



Locate your score.

My score of 600 is 1 standard deviation above the mean. That corresponds to one of the points of the 68–95–99.7 Rule.

TELL

Conclusion Interpret your result in context.

About 68% of those who took the test had scores that fell no more than 1 standard deviation from the mean, so 100% − 68% = 32% of all students had scores more than 1 standard deviation away. Only half of those were on the high side, so about 16% (half of 32%) of the test scores were better than mine. My score of 600 is higher than about 84% of all scores on this test.

The bounds of SAT scoring at 200 and 800 can also be explained by the 68–95–99.7 Rule. Since 200 and 800 are three standard deviations from 500, it hardly pays to extend the scoring any farther on either side. We'd get more information only on 100 − 99.7 = 0.3% of students.

The Worst-Case Scenario* Suppose we encounter an observation that's 5 standard deviations above the mean. Should we be surprised? We've just seen that when a Normal model is appropriate, such a value is exceptionally rare. After all, 99.7% of all the values should be within 3 standard deviations of the mean, so anything farther away would be unusual indeed.

But our handy 68–95–99.7 Rule applies only to Normal models, and the Normal is such a *nice* shape. What if we're dealing with a distribution that's strongly skewed (like the CEO salaries), or one that is uniform or bimodal or something really strange? A Normal model has 68% of its observations within one standard deviation of the mean, but a bimodal distribution could even be entirely empty in the middle. In that case could we still say anything at all about an observation 5 standard deviations above the mean?

Remarkably, even with really weird distributions, the worst case can't get all that bad. A Russian mathematician named Pafnuty Tchebycheff[5] answered the question by proving this theorem:

> *In any distribution, at least* $1 - \frac{1}{k^2}$ *of the values must lie within* $\pm k$ *standard deviations of the mean.*

What does that mean?

- For $k = 1$, $1 - \frac{1}{1^2} = 0$; if the distribution is far from Normal, it's possible that none of the values are within 1 standard deviation of the mean. We should be really cautious about saying anything about 68% unless we think a Normal model is justified. (Tchebycheff's theorem really is about the worst case; it tells us nothing about the middle; only about the extremes.)
- For $k = 2$, $1 - \frac{1}{2^2} = \frac{3}{4}$; no matter how strange the shape of the distribution, at least 75% of the values must be within 2 standard deviations of the mean. Normal models may expect 95% in that 2-standard-deviation interval, but even in a worst-case scenario it can never go lower than 75%.
- For $k = 3$, $1 - \frac{1}{3^2} = \frac{8}{9}$; in any distribution, at least 89% of the values lie within 3 standard deviations of the mean.

What we see is that values beyond 3 standard deviations from the mean are uncommon, Normal model or not. Tchebycheff tells us that at least 96% of all values must be within 5 standard deviations of the mean. While we can't always apply the 68–95–99.7 Rule, we can be sure that the observation we encountered 5 standard deviations above the mean is unusual.

Finding Normal Percentiles

A S ***Activity:* Your Pulse z-Score.** Is your pulse rate high or low? Find its *z*-score with the Normal Model Tool.

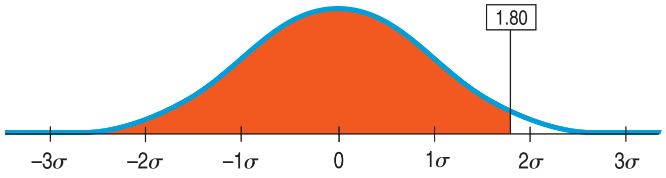
An SAT score of 600 is easy to assess, because we can think of it as one standard deviation above the mean. If your score was 680, though, where do you stand among the rest of the people tested? Your z-score is 1.80, so you're somewhere between 1 and 2 standard deviations above the mean. We figured out that no more than 16% of people score better than 600. By the same logic, no more than 2.5% of people score better than 700. Can we be more specific than "between 16% and 2.5%"?

When the value doesn't fall exactly 1, 2, or 3 standard deviations from the mean, we can look it up in a table of **Normal percentiles** or use technology.[6] Either way, we first convert our data to z-scores before using the table. Your SAT score of 680 has a z-score of $(680 - 500)/100 = 1.80$.

A S ***Activity:* The Normal Table.** Table Z just sits there, but this version of the Normal table changes so it always Makes a Picture that fits.

[5] He may have made the worst case for deviations clear, but the English spelling of his name is not. You'll find his first name spelled Pavnutii or Pavnuty and his last name spelled Chebsheff, Cebysev, and other creative versions.

[6] See Table Z in Appendix G, if you're curious. But your calculator (and any statistics computer package) does this, too—and more easily!



z	.00	.01
⋮	⋮	⋮
1.7	.9554	.9564
1.8	.9641	.9649
1.9	.9713	.9719
⋮	⋮	⋮

FIGURE 6.7

A table of Normal percentiles (Table Z in Appendix G) lets us find the percentage of individuals in a Standard Normal distribution falling below any specified z-score value.

TI-nspire

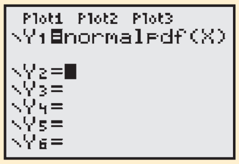
Normal percentiles. Explore the relationship between z-scores and areas in a Normal model.

In the piece of the table shown, we find your z-score by looking down the left column for the first two digits, 1.8, and across the top row for the third digit, 0. The table gives the percentile as 0.9641. That means that 96.4% of the z-scores are less than 1.80. Only 3.6% of people, then, scored better than 680 on the SAT.

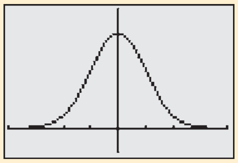
Most of the time, though, you'll do this with your calculator.

TI Tips

Finding Normal percentages



```
Plot1 Plot2 Plot3
\Y1■normalpdf(X)
\Y2=■
\Y3=
\Y4=
\Y5=
\Y6=
```

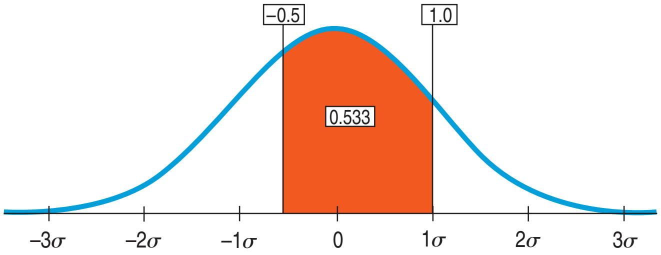


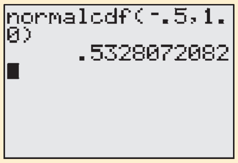
Your calculator knows the Normal model. Have a look under `2nd DISTR`. There you will see three "norm" functions, `normalpdf(`, `normalcdf(`, and `invNorm(`. Let's play with the first two.

- `normalpdf(` calculates y-values for graphing a Normal curve. You probably won't use this very often, if at all. If you want to try it, graph `Y1=normalpdf(X)` in a graphing `WINDOW` with `Xmin=-4`, `Xmax=4`, `Ymin=-0.1`, and `Ymax=0.5`.
- `normalcdf(` finds the proportion of area under the curve between two z-score cut points, by specifying `normalcdf(zLeft,zRight)`. Do make friends with this function; you will use it often!

Example 1

The Normal model shown shades the region between $z = -0.5$ and $z = 1.0$.





```
normalcdf(-.5,1.
0)
         .5328072082
■
```

To find the shaded area:

Under `2nd DISTR` select `normalcdf(`; hit `ENTER`.
Specify the cut points: `normalcdf(-.5,1.0)` and hit `ENTER` again.

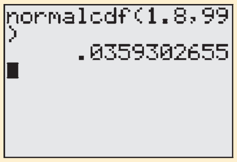
There's the area. Approximately 53% of a Normal model lies between half a standard deviation below and one standard deviation above the mean.

Example 2

In the example in the text we used Table Z to determine the fraction of SAT scores above your score of 680. Now let's do it again, this time using your `TI`.

First we need z-scores for the cut points:

- Since 680 is 1.8 standard deviations above the mean, your z-score is 1.8; that's the left cut point.



```
normalcdf(1.8,99
)
      .0359302655
▮
```

- Theoretically the standard Normal model extends rightward forever, but you can't tell the calculator to use infinity as the right cut point. Recall that for a Normal model almost all the area lies within ±3 standard deviations of the mean, so any upper cut point beyond, say, $z = 5$ does not cut off anything very important. We suggest you always use 99 (or −99) when you really want infinity as your cut point—it's easy to remember and way beyond any meaningful area.

Now you're ready. Use the command `normalcdf(1.8,99)`.

There you are! The Normal model estimates that approximately 3.6% of SAT scores are higher than 680.

STEP-BY-STEP EXAMPLE Working with Normal Models Part I

The Normal model is our first model for data. It's the first in a series of modeling situations where we step away from the data at hand to make more general statements about the world. We'll become more practiced in thinking about and learning the details of models as we progress through the book. To give you some practice in thinking about the Normal model, here are several problems that ask you to find percentiles in detail.

Question: What proportion of SAT scores fall between 450 and 600?

THINK

Plan State the problem.

I want to know the proportion of SAT scores between 450 and 600.

Variables Name the variable.

Let y = SAT score.

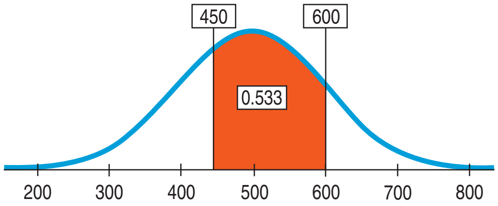
Check the appropriate conditions and specify which Normal model to use.

✓ **Nearly Normal Condition:** We are told that SAT scores are nearly Normal.

I'll model SAT scores with a N(500, 100) model, using the mean and standard deviation specified for them.

SHOW

Mechanics Make a picture of this Normal model. Locate the desired values and shade the region of interest.



Find z-scores for the cut points 450 and 600. Use technology to find the desired proportions, represented by the area under the curve. (This was Example 1 in the TI Tips—take another look.)

Standardizing the two scores, I find that

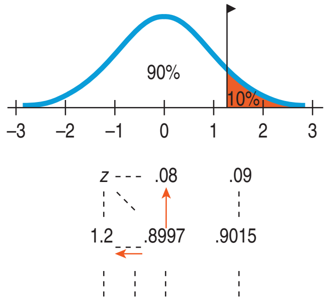
$$z = \frac{(y - \mu)}{\sigma} = \frac{(600 - 500)}{100} = 1.00$$

and

$$z = \frac{(450 - 500)}{100} = -0.50$$

	(If you use a table, then you need to subtract the two areas to find the area *between* the cut points.)	So, $\text{Area}(450 < y < 600) = \text{Area}(-0.5 < z < 1.0) = 0.5328$ (**OR**: From Table Z, the area $(z < 1.0) = 0.8413$ and area $(z < -0.5) = 0.3085$, so the proportion of z-scores *between* them is $0.8413 - 0.3085 = 0.5328$, or 53.28%.)
TELL	**Conclusion** Interpret your result in context.	The Normal model estimates that about 53.3% of SAT scores fall between 450 and 600.

From Percentiles to Scores: *z* in Reverse



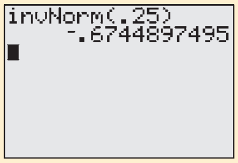
Finding areas from *z*-scores is the simplest way to work with the Normal model. But sometimes we start with areas and are asked to work backward to find the corresponding *z*-score or even the original data value. For instance, what *z*-score cuts off the top 10% in a Normal model?

Make a picture like the one shown, shading the rightmost 10% of the area. Notice that this is the 90th percentile. Look in Table Z for an area of 0.900. The exact area is not there, but 0.8997 is pretty close. That shows up in the table with 1.2 in the left margin and .08 in the top margin. The *z*-score for the 90th percentile, then, is approximately $z = 1.28$.

Computers and calculators will determine the cut point more precisely (and more easily).

TI Tips

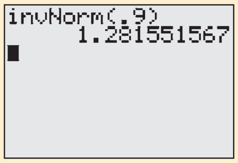
Finding Normal cutpoints



```
invNorm(.25)
    -.6744897495
```

To find the *z*-score at the 25th percentile, go to `2nd DISTR` again. This time we'll use the third of the "norm" functions, `invNorm(`.

Just specify the desired percentile with the command `invNorm(.25)` and hit `ENTER`. The calculator says that the cut point for the leftmost 25% of a Normal model is approximately $z = -0.674$.



```
invNorm(.9)
     1.281551567
```

One more example: What *z*-score cuts off the highest 10% of a Normal model? That's easily done—just remember to specify the *percentile*. Since we want the cut point for the *highest* 10%, we know that the other 90% must be *below* that *z*-score. The cut point, then, must stand at the 90th percentile, so specify `invNorm(.90)`.

Only 10% of the area in a Normal model is more than about 1.28 standard deviations above the mean.

STEP-BY-STEP EXAMPLE Working with Normal Models Part II

Question: Suppose a college says it admits only people with SAT Verbal test scores among the top 10%. How high a score does it take to be eligible?

THINK

Plan State the problem.

How high an SAT Verbal score do I need to be in the top 10% of all test takers?

Variable Define the variable.

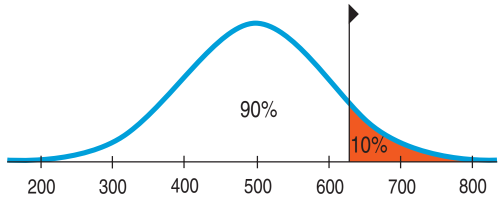
Let y = my SAT score.

Check to see if a Normal model is appropriate, and specify which Normal model to use.

✓ **Nearly Normal Condition:** I am told that SAT scores are nearly Normal. I'll model them with *N*(500, 100).

SHOW

Mechanics Make a picture of this Normal model. Locate the desired percentile approximately by shading the rightmost 10% of the area.



The college takes the top 10%, so its cutoff score is the 90th percentile. Find the corresponding *z*-score using your calculator as shown in the TI Tips. (**OR**: Use Table Z as shown on p. 119.)

The cut point is $z = 1.28$.

Convert the *z*-score back to the original units.

A *z*-score of 1.28 is 1.28 standard deviations above the mean. Since the SD is 100, that's 128 SAT points. The cutoff is 128 points above the mean of 500, or 628.

TELL

Conclusion Interpret your results in the proper context.

Because the school wants SAT Verbal scores in the top 10%, the cutoff is 628. (Actually, since SAT scores are reported only in multiples of 10, I'd have to score at least a 630.)

TI-*nspire*

Normal models. Watch the Normal model react as you change the mean and standard deviation.

STEP-BY-STEP EXAMPLE More Working with Normal Models

Working with Normal percentiles can be a little tricky, depending on how the problem is stated. Here are a few more worked examples of the kind you're likely to see.

A cereal manufacturer has a machine that fills the boxes. Boxes are labeled "16 ounces," so the company wants to have that much cereal in each box, but since no packaging process is perfect, there will be minor variations. If the machine is set at exactly 16 ounces and the Normal model applies (or at least the distribution is roughly symmetric), then about half of the boxes will be underweight, making consumers unhappy and exposing the company to bad publicity and possible lawsuits. To prevent underweight boxes, the manufacturer has to set the mean a little higher than 16.0 ounces.

Based on their experience with the packaging machine, the company believes that the amount of cereal in the boxes fits a Normal model with a standard deviation of 0.2 ounces. The manufacturer decides to set the machine to put an average of 16.3 ounces in each box. Let's use that model to answer a series of questions about these cereal boxes.

Question 1: What fraction of the boxes will be underweight?

THINK

Plan State the problem.

What proportion of boxes weigh less than 16 ounces?

Variable Name the variable.

Let y = weight of cereal in a box.

Check to see if a Normal model is appropriate.

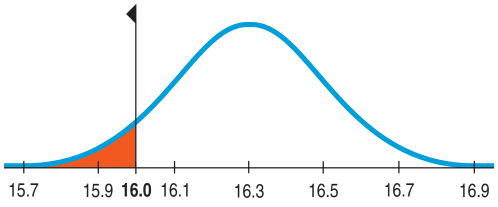
✓ **Nearly Normal Condition:** I have no data, so I cannot make a histogram, but I am told that the company believes the distribution of weights from the machine is Normal.

Specify which Normal model to use.

I'll use a $N(16.3, 0.2)$ model.

SHOW

Mechanics Make a picture of this Normal model. Locate the value you're interested in on the picture, label it, and shade the appropriate region.



REALITY CHECK

Estimate from the picture the percentage of boxes that are underweight. (This will be useful later to check that your answer makes sense.) It looks like a low percentage. Less than 20% for sure.

I want to know what fraction of the boxes will weigh less than 16 ounces.

Convert your cutoff value into a z-score.

$$z = \frac{y - \mu}{\sigma} = \frac{16 - 16.3}{0.2} = -1.50$$

Find the area with your calculator (or use the Normal table).

$$\text{Area}\ (y < 16) = \text{Area}(z < -1.50) = 0.0668$$

TELL

Conclusion State your conclusion, and check that it's consistent with your earlier guess. It's below 20%—seems okay.

I estimate that approximately 6.7% of the boxes will contain less than 16 ounces of cereal.

Question 2: The company's lawyers say that 6.7% is too high. They insist that no more than 4% of the boxes can be underweight. So the company needs to set the machine to put a little more cereal in each box. What mean setting do they need?

THINK

Plan State the problem.

What mean weight will reduce the proportion of underweight boxes to 4%?

Variable Name the variable.

Let y = weight of cereal in a box.

Check to see if a Normal model is appropriate.

✓ **Nearly Normal Condition:** I am told that a Normal model applies.

Specify which Normal model to use. This time you are not given a value for the mean!

I don't know μ, the mean amount of cereal. The standard deviation for this machine is 0.2 ounces. The model is $N(\mu, 0.2)$.

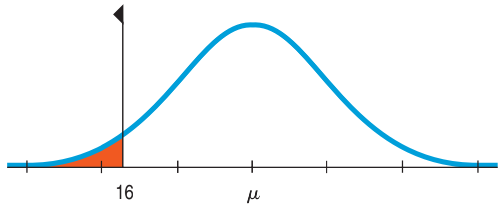
REALITY CHECK

We found out earlier that setting the machine to $\mu = 16.3$ ounces made 6.7% of the boxes too light. We'll need to raise the mean a bit to reduce this fraction.

No more than 4% of the boxes can be below 16 ounces.

SHOW

Mechanics Make a picture of this Normal model. Center it at μ (since you don't know the mean), and shade the region below 16 ounces.



Using your calculator (or the Normal table), find the z-score that cuts off the lowest 4%.

The z-score that has 0.04 area to the left of it is $z = -1.75$.

Use this information to find μ. It's located 1.75 standard deviations to the right of 16. Since σ is 0.2, that's 1.75 × 0.2, or 0.35 ounces more than 16.

For 16 to be 1.75 standard deviations below the mean, the mean must be

$$16 + 1.75\,(0.2) = 16.35 \text{ ounces.}$$

TELL

Conclusion Interpret your result in context.
(This makes sense; we knew it would have to be just a bit higher than 16.3.)

The company must set the machine to average 16.35 ounces of cereal per box.

Question 3: The company president vetoes that plan, saying the company should give away less free cereal, not more. Her goal is to set the machine no higher than 16.2 ounces and still have only 4% underweight boxes. The only way to accomplish this is to reduce the standard deviation. What standard deviation must the company achieve, and what does that mean about the machine?

THINK

Plan State the problem.

What standard deviation will allow the mean to be 16.2 ounces and still have only 4% of boxes underweight?

Variable Name the variable.

Let y = weight of cereal in a box.

Check conditions to be sure that a Normal model is appropriate.

✓ **Nearly Normal Condition:** The company believes that the weights are described by a Normal model.

Specify which Normal model to use. This time you don't know σ.

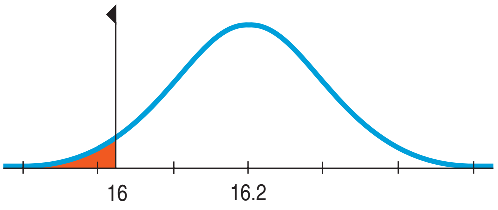
I know the mean, but not the standard deviation, so my model is $N(16.2, \sigma)$.

REALITY CHECK

We know the new standard deviation must be less than 0.2 ounces.

SHOW

Mechanics Make a picture of this Normal model. Center it at 16.2, and shade the area you're interested in. We want 4% of the area to the left of 16 ounces.



Find the z-score that cuts off the lowest 4%.

I know that the z-score with 4% below it is $z = -1.75$.

Solve for σ. (We need 16 to be 1.75 σ's below 16.2, so 1.75 σ must be 0.2 ounces. You could just start with that equation.)

$$
\begin{aligned}
z &= \frac{y - \mu}{\sigma} \\
-1.75 &= \frac{16 - 16.2}{\sigma} \\
1.75\,\sigma &= 0.2 \\
\sigma &= 0.114
\end{aligned}
$$

TELL

Conclusion Interpret your result in context.

As we expected, the standard deviation is lower than before—actually, quite a bit lower.

The company must get the machine to box cereal with a standard deviation of only 0.114 ounces. This means the machine must be more consistent (by nearly a factor of 2) in filling the boxes.

Are You Normal? Find Out with a Normal Probability Plot

In the examples we've worked through, we've assumed that the underlying data distribution was roughly unimodal and symmetric, so that using a Normal model makes sense. When you have data, you must *check* to see whether a Normal model is reasonable. How? Make a picture, of course! Drawing a histogram of the data and looking at the shape is one good way to see if a Normal model might work.

TI-*nspire*

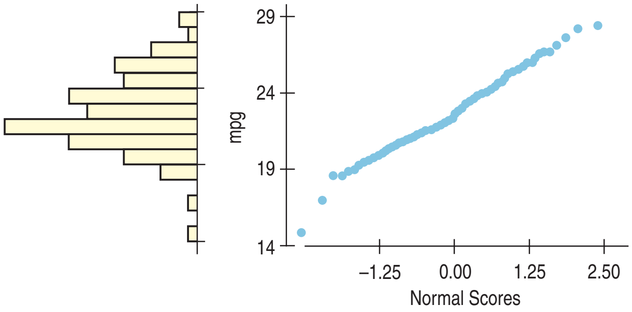
Normal probability plots and histograms. See how a normal probability plot responds as you change the shape of a distribution.

There's a more specialized graphical display that can help you to decide whether the Normal model is appropriate: the **Normal probability plot.** If the distribution of the data is roughly Normal, the plot is roughly a diagonal straight line. Deviations from a straight line indicate that the distribution is not Normal. This plot is usually able to show deviations from Normality more clearly than the corresponding histogram, but it's usually easier to understand *how* a distribution fails to be Normal by looking at its histogram.

Some data on a car's fuel efficiency provide an example of data that are nearly Normal. The overall pattern of the Normal probability plot is straight. The two trailing low values correspond to the values in the histogram that trail off the low end. They're not quite in line with the rest of the data set. The Normal probability plot shows us that they're a bit lower than we'd expect of the lowest two values in a Normal model.

FIGURE 6.9

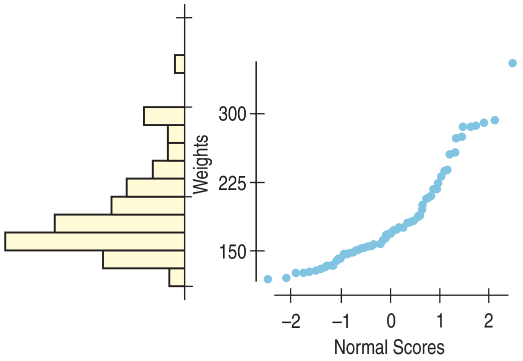
Histogram and Normal probability plot for gas mileage (mpg) recorded by one of the authors over the 8 years he owned a 1989 Nissan Maxima. The vertical axes are the same, so each dot on the probability plot would fall into the bar on the histogram immediately to its left.



By contrast, the Normal probability plot of the men's *Weights* from the NHANES Study is far from straight. The weights are skewed to the high end, and the plot is curved. We'd conclude from these pictures that approximations using the 68–95–99.7 Rule for these data would not be very accurate.

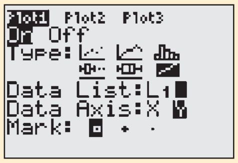
FIGURE 6.10

Histogram and Normal probability plot for men's weights. Note how a skewed distribution corresponds to a bent probability plot.

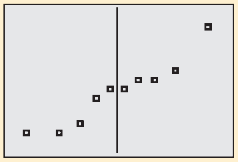


TI Tips

Creating a Normal probability plot



```
Plot1 Plot2 Plot3
On Off
Type:
Data List:L1
Data Axis:X Y
Mark: □ + ·
```



Let's make a Normal probability plot with the calculator. Here are the boys' agility test scores we looked at in Chapter 5; enter them in L1:

22, 17, 18, 29, 22, 23, 24, 23, 17, 21

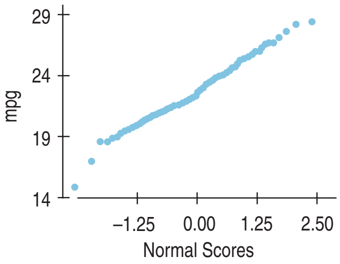
Now you can create the plot:

- Turn a STATPLOT On.
- Tell it to make a Normal probability plot by choosing the last of the icons.
- Specify your datalist and which axis you want the data on. (We'll use Y so the plot looks like the others we showed you.)
- Specify the Mark you want the plot to use.
- Now ZoomStat does the rest.

The plot doesn't look very straight. Normality is certainly questionable here.

(Not that it matters in making this decision, but that vertical line is the y-axis. Points to the left have negative z-scores and points to the right have positive z-scores.)

How Does a Normal Probability Plot Work?



Why does the Normal probability plot work like that? We looked at 100 fuel efficiency measures for the author's Nissan car. The smallest of these has a z-score of -3.16. The Normal model can tell us what value to expect for the smallest z-score in a batch of 100 if a Normal model were appropriate. That turns out to be -2.58. So our first data value is smaller than we would expect from the Normal.

We can continue this and ask a similar question for each value. For example, the 14th-smallest fuel efficiency has a z-score of almost exactly -1, and that's just what we should expect (well, -1.1 to be exact). A Normal probability plot takes each data value and plots it against the z-score you'd expect that point to have if the distribution were perfectly Normal.[7]

When the values match up well, the line is straight. If one or two points are surprising from the Normal's point of view, they don't line up. When the entire distribution is skewed or different from the Normal in some other way, the values don't match up very well at all and the plot bends.

It turns out to be tricky to find the values we expect. They're called *Normal scores*, but you can't easily look them up in the tables. That's why probability plots are best made with technology and not by hand.

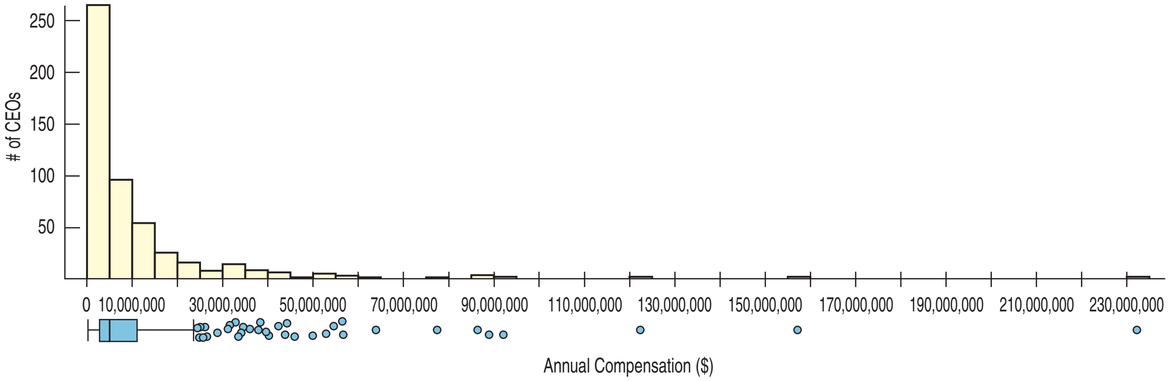
The best advice on using Normal probability plots is to see whether they are straight. If so, then your data look like data from a Normal model. If not, make a histogram to understand how they differ from the model.

A S *Activity:* **Assessing Normality.** This activity guides you through the process of checking the Nearly Normal condition using your statistics package.

[7] Sometimes the Normal probability plot switches the two axes, putting the data on the x-axis and the z-scores on the y-axis.

WHAT CAN GO WRONG?

- **Don't use a Normal model when the distribution is not unimodal and symmetric.** Normal models are so easy and useful that it is tempting to use them even when they don't describe the data very well. That can lead to wrong conclusions. Don't use a Normal model without first checking the **Nearly Normal Condition.** Look at a picture of the data to check that it is unimodal and symmetric. A histogram, or a Normal probability plot, can help you tell whether a Normal model is appropriate.



The CEOs (p. 90) had a mean total compensation of $10,307,311.87 with a standard deviation of $17,964,615.16. Using the Normal model rule, we should expect about 68% of the CEOs to have compensations between −$7,657,303.29 and $28,271,927.03. In fact, more than 90% of the CEOs have annual compensations in this range. What went wrong? The distribution is skewed, not symmetric. Using the 68–95–99.7 Rule for data like these will lead to silly results.

- **Don't use the mean and standard deviation when outliers are present.** Both means and standard deviations can be distorted by outliers, and no model based on distorted values will do a good job. A z-score calculated from a distribution with outliers may be misleading. It's always a good idea to check for outliers. How? Make a picture.
- **Don't round your results in the middle of a calculation.** We *reported* the mean of the heptathletes' long jump as 6.16 meters. More precisely, it was 6.16153846153846 meters.

  You should use all the precision available in the data for all the intermediate steps of a calculation. Using the more precise value for the mean (and also carrying 15 digits for the SD), the z-score calculation for Klüft's long jump comes out to

$$z = \frac{6.78 - 6.16153846153846}{0.2297597407326585} = 2.691775053755667700$$

  We'd report that as 2.692, as opposed to the rounded-off value of 2.70 we got earlier from the table.
- **Don't worry about minor differences in results.** Because various calculators and programs may carry different precision in calculations, your answers may differ slightly from those we show in the text and in the Step-By-Steps, or even from the values given in the answers in the back of the book. Those differences aren't anything to worry about. They're not the main story Statistics tries to tell.

CONNECTIONS

Changing the center and spread of a variable is equivalent to changing its *units.* Indeed, the only part of the data's context changed by standardizing is the units. All other aspects of the context do not depend on the choice or modification of measurement units. This fact points out an important distinction between the numbers the data provide for calculation and the meaning of the variables and the relationships among them. Standardizing can make the numbers easier to work with, but it does not alter the meaning.

Another way to look at this is to note that standardizing may change the center and spread values, but it does not affect the *shape* of a distribution. A histogram or boxplot of standardized values looks just the same as the histogram or boxplot of the original values except, perhaps, for the numbers on the axes.

When we summarized *shape, center,* and *spread* for histograms, we compared them to unimodal, symmetric shapes. You couldn't ask for a nicer example than the Normal model. And if the shape *is* like a Normal, we'll use the the mean and standard deviation to standardize the values.

WHAT HAVE WE LEARNED?

We've learned that the story data can tell may be easier to understand after shifting or rescaling the data.

- Shifting data by adding or subtracting the same amount from each value affects measures of center and position but not measures of spread.
- Rescaling data by multiplying or dividing every value by a constant, changes all the summary statistics—center, position, and spread.

We've learned the power of standardizing data.

- Standardizing uses the standard deviation as a ruler to measure distance from the mean, creating *z*-scores.
- Using these *z*-scores, we can compare apples and oranges—values from different distributions or values based on different units.
- And a *z*-score can identify unusual or surprising values among data.

We've learned that the 68–95–99.7 Rule can be a useful rule of thumb for understanding distributions.

- For data that are unimodal and symmetric, about 68% fall within 1 SD of the mean, 95% fall within 2 SDs of the mean, and 99.7% fall within 3 SDs of the mean (see p. 130).

Again we've seen the importance of *Thinking* about whether a method will work.

- **Normality Assumption:** We sometimes work with Normal tables (Table Z). Those tables are based on the Normal model.
- Data can't be exactly Normal, so we check the **Nearly Normal Condition** by making a histogram (is it unimodal, symmetric, and free of outliers?) or a Normal probability plot (is it straight enough?). (See p. 125.)

Terms

Standardizing

105. We standardize to eliminate units. Standardized values can be compared and combined even if the original variables had different units and magnitudes.

Standardized value

105. A value found by subtracting the mean and dividing by the standard deviation.

Shifting 107. Adding a constant to each data value adds the same constant to the mean, the median, and the quartiles, but does not change the standard deviation or IQR.

Rescaling 108. Multiplying each data value by a constant multiplies both the measures of position (mean, median, and quartiles) and the measures of spread (standard deviation and IQR) by that constant.

Normal model 112. A useful family of models for unimodal, symmetric distributions.

Parameter 112. A numerically valued attribute of a model. For example, the values of μ and σ in a $N(\mu, \sigma)$ model are parameters.

Statistic 112. A value calculated from data to summarize aspects of the data. For example, the mean, $\bar{y}$ and standard deviation, s, are statistics.

***z*-score** 105. A *z*-score tells how many standard deviations a value is from the mean; *z*-scores have a mean of 0 and a standard deviation of 1. When working with data, use the statistics $\bar{y}$ and s:

$$z = \frac{y - \bar{y}}{s}.$$

112. When working with models, use the parameters μ and σ:

$$z = \frac{y - \mu}{\sigma}.$$

Standard Normal model 112. A Normal model, $N(\mu, \sigma)$ with mean $\mu = 0$ and standard deviation $\sigma = 1$. Also called the **standard Normal distribution.**

Nearly Normal Condition 112. A distribution is nearly Normal if it is unimodal and symmetric. We can check by looking at a histogram or a Normal probability plot.

68–95–99.7 Rule 113. In a Normal model, about 68% of values fall within 1 standard deviation of the mean, about 95% fall within 2 standard deviations of the mean, and about 99.7% fall within 3 standard deviations of the mean.

Normal percentile 116. The Normal percentile corresponding to a *z*-score gives the percentage of values in a standard Normal distribution found at that *z*-score or below.

Normal probability plot 124. A display to help assess whether a distribution of data is approximately Normal. If the plot is nearly straight, the data satisfy the **Nearly Normal Condition.**

Skills

THINK

- Understand how adding (subtracting) a constant or multiplying (dividing) by a constant changes the center and/or spread of a variable.
- Recognize when standardization can be used to compare values.
- Understand that standardizing uses the standard deviation as a ruler.
- Recognize when a Normal model is appropriate.

SHOW

- Know how to calculate the *z*-score of an observation.
- Know how to compare values of two different variables using their *z*-scores.
- Be able to use Normal models and the 68–95–99.7 Rule to estimate the percentage of observations falling within 1, 2, or 3 standard deviations of the mean.
- Know how to find the percentage of observations falling below any value in a Normal model using a Normal table or appropriate technology.
- Know how to check whether a variable satisfies the **Nearly Normal Condition** by making a Normal probability plot or a histogram.

TELL

- Know what *z*-scores mean.
- Be able to explain how extraordinary a standardized value may be by using a Normal model.

NORMAL PLOTS ON THE COMPUTER

The best way to tell whether your data can be modeled well by a Normal model is to make a picture or two. We've already talked about making histograms. Normal probability plots are almost never made by hand because the values of the Normal scores are tricky to find. But most statistics software make Normal plots, though various packages call the same plot by different names and array the information differently.

EXERCISES

1. **Shipments.** A company selling clothing on the Internet reports that the packages it ships have a median weight of 68 ounces and an IQR of 40 ounces.
   a) The company plans to include a sales flyer weighing 4 ounces in each package. What will the new median and IQR be?
   b) If the company recorded the shipping weights of these new packages in pounds instead of ounces, what would the median and IQR be? (1 lb. = 16 oz.)

2. **Hotline.** A company's customer service hotline handles many calls relating to orders, refunds, and other issues. The company's records indicate that the median length of calls to the hotline is 4.4 minutes with an IQR of 2.3 minutes.
   a) If the company were to describe the duration of these calls in seconds instead of minutes, what would the median and IQR be?
   b) In an effort to speed up the customer service process, the company decides to streamline the series of push-button menus customers must navigate, cutting the time by 24 seconds. What will the median and IQR of the length of hotline calls become?

3. **Payroll.** Here are the summary statistics for the weekly payroll of a small company: lowest salary = \$300, mean salary = \$700, median = \$500, range = \$1200, IQR = \$600, first quartile = \$350, standard deviation = \$400.
   a) Do you think the distribution of salaries is symmetric, skewed to the left, or skewed to the right? Explain why.
   b) Between what two values are the middle 50% of the salaries found?
   c) Suppose business has been good and the company gives every employee a \$50 raise. Tell the new value of each of the summary statistics.
   d) Instead, suppose the company gives each employee a 10% raise. Tell the new value of each of the summary statistics.

4. **Hams.** A specialty foods company sells "gourmet hams" by mail order. The hams vary in size from 4.15 to 7.45 pounds, with a mean weight of 6 pounds and standard deviation of 0.65 pounds. The quartiles and median weights are 5.6, 6.2, and 6.55 pounds.
   a) Find the range and the IQR of the weights.
   b) Do you think the distribution of the weights is symmetric or skewed? If skewed, which way? Why?
   c) If these weights were expressed in ounces (1 pound = 16 ounces) what would the mean, standard deviation, quartiles, median, IQR, and range be?
   d) When the company ships these hams, the box and packing materials add 30 ounces. What are the mean, standard deviation, quartiles, median, IQR, and range of weights of boxes shipped (in ounces)?
   e) One customer made a special order of a 10-pound ham. Which of the summary statistics of part d might *not* change if that data value were added to the distribution?

5. **SAT or ACT?** Each year thousands of high school students take either the SAT or the ACT, standardized tests used in the college admissions process. Combined SAT Math and Verbal scores go as high as 1600, while the maximum ACT composite score is 36. Since the two exams use very different scales, comparisons of performance are difficult. A convenient rule of thumb is $SAT = 40 \times ACT + 150$; that is, multiply an ACT score by 40 and add 150 points to estimate the equivalent SAT score. An admissions officer reported the following statistics about the ACT scores of 2355 students who applied to her college one year. Find the summaries of equivalent SAT scores.

   Lowest score = 19 Mean = 27 Standard deviation = 3

   Q3 = 30 Median = 28 IQR = 6

6. **Cold U?** A high school senior uses the Internet to get information on February temperatures in the town where he'll be going to college. He finds a Web site with some statistics, but they are given in degrees Celsius. The conversion formula is $^\circ F = 9/5\ ^\circ C + 32$. Determine the Fahrenheit equivalents for the summary information below.

   Maximum temperature = 11°C Range = 33°

   Mean = 1° Standard deviation = 7°

   Median = 2° IQR = 16°

7. **Stats test.** Suppose your Statistics professor reports test grades as z-scores, and you got a score of 2.20 on an exam. Write a sentence explaining what that means.

8. **Checkup.** One of the authors has an adopted grandson whose birth family members are very short. After examining him at his 2-year checkup, the boy's pediatrician said that the z-score for his height relative to American 2-year-olds was -1.88. Write a sentence explaining what that means.

9. **Stats test, part II.** The mean score on the Stats exam was 75 points with a standard deviation of 5 points, and Gregor's z-score was -2. How many points did he score?

10. **Mensa.** People with z-scores above 2.5 on an IQ test are sometimes classified as geniuses. If IQ scores have a mean of 100 and a standard deviation of 16 points, what IQ score do you need to be considered a genius?

11. **Temperatures.** A town's January high temperatures average 36°F with a standard deviation of 10°, while in July the mean high temperature is 74° and the standard deviation is 8°. In which month is it more unusual to have a day with a high temperature of 55°? Explain.

12. **Placement exams.** An incoming freshman took her college's placement exams in French and mathematics. In French, she scored 82 and in math 86. The overall results on the French exam had a mean of 72 and a standard deviation of 8, while the mean math score was 68, with a standard deviation of 12. On which exam did she do better compared with the other freshmen?

13. **Combining test scores.** The first Stats exam had a mean of 65 and a standard deviation of 10 points; the second had a mean of 80 and a standard deviation of 5 points. Derrick scored an 80 on both tests. Julie scored a 70 on the first test and a 90 on the second. They both totaled 160 points on the two exams, but Julie claims that her total is better. Explain.

14. **Combining scores again.** The first Stat exam had a mean of 80 and a standard deviation of 4 points; the second had a mean of 70 and a standard deviation of 15 points. Reginald scored an 80 on the first test and an 85 on the second. Sara scored an 88 on the first but only a 65 on the second. Although Reginald's total score is higher, Sara feels she should get the higher grade. Explain her point of view.

15. **Final exams.** Anna, a language major, took final exams in both French and Spanish and scored 83 on each. Her roommate Megan, also taking both courses, scored 77 on the French exam and 95 on the Spanish exam. Overall, student scores on the French exam had a mean of 81 and a standard deviation of 5, and the Spanish scores had a mean of 74 and a standard deviation of 15.
    a) To qualify for language honors, a major must maintain at least an 85 average for all language courses taken. So far, which student qualifies?
    b) Which student's overall performance was better?

16. **MP3s.** Two companies market new batteries targeted at owners of personal music players. DuraTunes claims a mean battery life of 11 hours, while RockReady advertises 12 hours.
    a) Explain why you would also like to know the standard deviations of the battery lifespans before deciding which brand to buy.
    b) Suppose those standard deviations are 2 hours for DuraTunes and 1.5 hours for RockReady. You are headed for 8 hours at the beach. Which battery is most likely to last all day? Explain.
    c) If your beach trip is all weekend, and you probably will have the music on for 16 hours, which battery is most likely to last? Explain.

17. **Cattle.** The Virginia Cooperative Extension reports that the mean weight of yearling Angus steers is 1152 pounds. Suppose that weights of all such animals can be described by a Normal model with a standard deviation of 84 pounds.
    a) How many standard deviations from the mean would a steer weighing 1000 pounds be?
    b) Which would be more unusual, a steer weighing 1000 pounds or one weighing 1250 pounds?

T 18. **Car speeds.** John Beale of Stanford, CA, recorded the speeds of cars driving past his house, where the speed limit read 20 mph. The mean of 100 readings was 23.84 mph, with a standard deviation of 3.56 mph. (He actually recorded every car for a two-month period. These are 100 representative readings.)
    a) How many standard deviations from the mean would a car going under the speed limit be?
    b) Which would be more unusual, a car traveling 34 mph or one going 10 mph?

19. **More cattle.** Recall that the beef cattle described in Exercise 17 had a mean weight of 1152 pounds, with a standard deviation of 84 pounds.
    a) Cattle buyers hope that yearling Angus steers will weigh at least 1000 pounds. To see how much over (or under) that goal the cattle are, we could subtract 1000 pounds from all the weights. What would the new mean and standard deviation be?
    b) Suppose such cattle sell at auction for 40 cents a pound. Find the mean and standard deviation of the sale prices for all the steers.

T 20. **Car speeds again.** For the car speed data of Exercise 18, recall that the mean speed recorded was 23.84 mph, with a standard deviation of 3.56 mph. To see how many cars are speeding, John subtracts 20 mph from all speeds.
    a) What is the mean speed now? What is the new standard deviation?
    b) His friend in Berlin wants to study the speeds, so John converts all the original miles-per-hour readings to kilometers per hour by multiplying all speeds by 1.609 (km per mile). What is the mean now? What is the new standard deviation?

21. **Cattle, part III.** Suppose the auctioneer in Exercise 19 sold a herd of cattle whose minimum weight was 980 pounds, median was 1140 pounds, standard deviation 84 pounds, and IQR 102 pounds. They sold for 40 cents a pound, and the auctioneer took a $20 commission on each animal. Then, for example, a steer weighing 1100 pounds would net the owner 0.40 (1100) − 20 = $420. Find the minimum, median, standard deviation, and IQR of the net sale prices.

22. **Caught speeding.** Suppose police set up radar surveillance on the Stanford street described in Exercise 18. They handed out a large number of tickets to speeders going a mean of 28 mph, with a standard deviation of 2.4 mph, a maximum of 33 mph, and an IQR of 3.2 mph. Local law prescribes fines of $100, plus $10 per mile per hour over the 20 mph speed limit. For example, a driver convicted of going 25 mph would be fined 100 + 10(5) = $150. Find the mean, maximum, standard deviation, and IQR of all the potential fines.

23. **Professors.** A friend tells you about a recent study dealing with the number of years of teaching experience among current college professors. He remembers the mean but can't recall whether the standard deviation was 6 months, 6 years, or 16 years. Tell him which one it must have been, and why.

24. **Rock concerts.** A popular band on tour played a series of concerts in large venues. They always drew a large crowd, averaging 21,359 fans. While the band did not announce (and probably never calculated) the standard deviation, which of these values do you think is most likely to be correct: 20, 200, 2000, or 20,000 fans? Explain your choice.

25. **Guzzlers?** Environmental Protection Agency (EPA) fuel economy estimates for automobile models tested recently predicted a mean of 24.8 mpg and a standard deviation of 6.2 mpg for highway driving. Assume that a Normal model can be applied.
a) Draw the model for auto fuel economy. Clearly label it, showing what the 68–95–99.7 Rule predicts.
b) In what interval would you expect the central 68% of autos to be found?
c) About what percent of autos should get more than 31 mpg?
d) About what percent of cars should get between 31 and 37.2 mpg?
e) Describe the gas mileage of the worst 2.5% of all cars.

26. **IQ.** Some IQ tests are standardized to a Normal model, with a mean of 100 and a standard deviation of 16.
a) Draw the model for these IQ scores. Clearly label it, showing what the 68–95–99.7 Rule predicts.
b) In what interval would you expect the central 95% of IQ scores to be found?
c) About what percent of people should have IQ scores above 116?
d) About what percent of people should have IQ scores between 68 and 84?
e) About what percent of people should have IQ scores above 132?

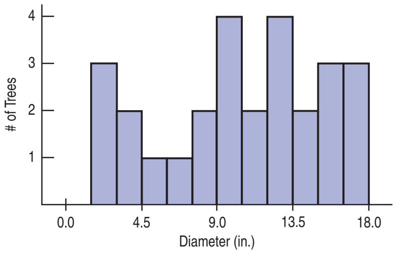
27. **Small steer.** In Exercise 17 we suggested the model *N*(1152, 84) for weights in pounds of yearling Angus steers. What weight would you consider to be unusually low for such an animal? Explain.

28. **High IQ.** Exercise 26 proposes modeling IQ scores with *N*(100, 16). What IQ would you consider to be unusually high? Explain.

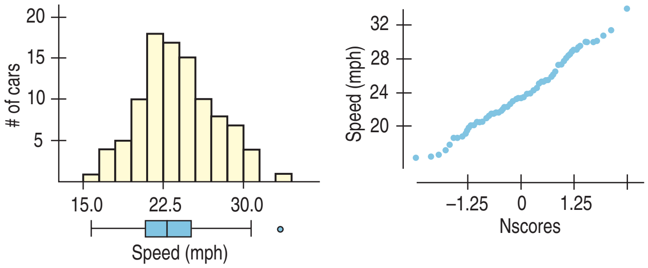
29. **Trees.** A forester measured 27 of the trees in a large woods that is up for sale. He found a mean diameter of 10.4 inches and a standard deviation of 4.7 inches. Suppose that these trees provide an accurate description of the whole forest and that a Normal model applies.
a) Draw the Normal model for tree diameters.
b) What size would you expect the central 95% of all trees to be?
c) About what percent of the trees should be less than an inch in diameter?
d) About what percent of the trees should be between 5.7 and 10.4 inches in diameter?
e) About what percent of the trees should be over 15 inches in diameter?

30. **Rivets.** A company that manufactures rivets believes the shear strength (in pounds) is modeled by *N*(800, 50).
a) Draw and label the Normal model.
b) Would it be safe to use these rivets in a situation requiring a shear strength of 750 pounds? Explain.
c) About what percent of these rivets would you expect to fall below 900 pounds?
d) Rivets are used in a variety of applications with varying shear strength requirements. What is the maximum shear strength for which you would feel comfortable approving this company's rivets? Explain your reasoning.

31. **Trees, part II.** Later on, the forester in Exercise 29 shows you a histogram of the tree diameters he used in analyzing the woods that was for sale. Do you think he was justified in using a Normal model? Explain, citing some specific concerns.



T 32. **Car speeds, the picture.** For the car speed data of Exercise 18, here is the histogram, boxplot, and Normal probability plot of the 100 readings. Do you think it is appropriate to apply a Normal model here? Explain.



T 33. **Winter Olympics 2006 downhill.** Fifty-three men qualified for the men's alpine downhill race in Torino. The gold medal winner finished in 1 minute, 48.8 seconds. All competitors' times (in seconds) are found in the following list:

108.80	109.52	109.82	109.88	109.93	110.00
110.04	110.12	110.29	110.33	110.35	110.44
110.45	110.64	110.68	110.70	110.72	110.84
110.88	110.88	110.90	110.91	110.98	111.37
111.48	111.51	111.55	111.70	111.72	111.93
112.17	112.55	112.87	112.90	113.34	114.07
114.65	114.70	115.01	115.03	115.73	116.10
116.58	116.81	117.45	117.54	117.56	117.69
118.77	119.24	119.41	119.79	120.93	

a) The mean time was 113.02 seconds, with a standard deviation of 3.24 seconds. If the Normal model is appropriate, what percent of times will be less than 109.78 seconds?
b) What is the actual percent of times less than 109.78 seconds?
c) Why do you think the two percentages don't agree?
d) Create a histogram of these times. What do you see?

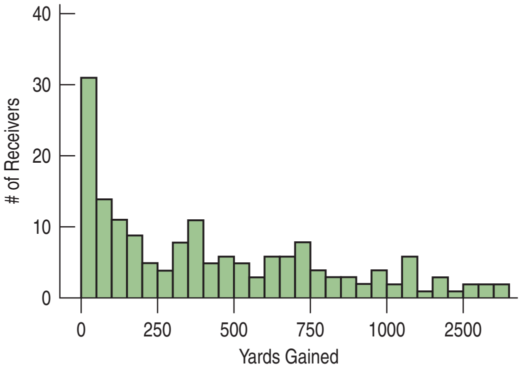
34. Check the model. The mean of the 100 car speeds in Exercise 20 was 23.84 mph, with a standard deviation of 3.56 mph.

a) Using a Normal model, what values should border the middle 95% of all car speeds?
b) Here are some summary statistics.

Percentile		Speed
100%	**Max**	34.060
97.5%		30.976
90.0%		28.978
75.0%	**Q3**	25.785
50.0%	**Median**	23.525
25.0%	**Q1**	21.547
10.0%		19.163
2.5%		16.638
0.0%	**Min**	16.270

From your answer in part a, how well does the model do in predicting those percentiles? Are you surprised? Explain.

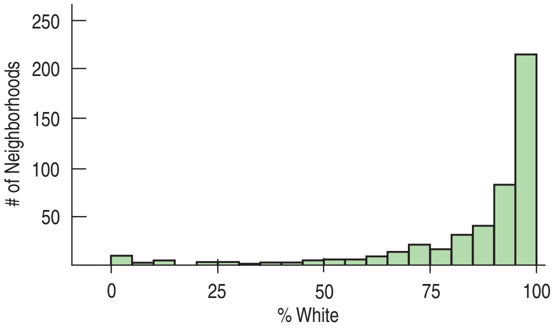
35. Receivers. NFL data from the 2006 football season reported the number of yards gained by each of the league's 167 wide receivers:



The mean is 435 yards, with a standard deviation of 384 yards.

a) According to the Normal model, what percent of receivers would you expect to gain fewer yards than 2 standard deviations below the mean number of yards?
b) For these data, what does that mean?
c) Explain the problem in using a Normal model here.

36. Customer database. A large philanthropic organization keeps records on the people who have contributed to their cause. In addition to keeping records of past giving, the organization buys demographic data on neighborhoods from the U.S. Census Bureau. Eighteen of these variables concern the ethnicity of the neighborhood of the donor. Here are a histogram and summary statistics for the percentage of whites in the neighborhoods of 500 donors:



```
Count     500
Mean       83.59
Median     93
StdDev     22.26
IQR        17
Q1         80
Q3         97
```

a) Which is a better summary of the percentage of white residents in the neighborhoods, the mean or the median? Explain.
b) Which is a better summary of the spread, the IQR or the standard deviation? Explain.
c) From a Normal model, about what percentage of neighborhoods should have a percent white within one standard deviation of the mean?
d) What percentage of neighborhoods actually have a percent white within one standard deviation of the mean?
e) Explain the discrepancy between parts c and d.

37. Normal cattle. Using $N(1152, 84)$, the Normal model for weights of Angus steers in Exercise 17, what percent of steers weigh

a) over 1250 pounds?
b) under 1200 pounds?
c) between 1000 and 1100 pounds?

38. IQs revisited. Based on the Normal model $N(100, 16)$ describing IQ scores, what percent of people's IQs would you expect to be

a) over 80?
b) under 90?
c) between 112 and 132?

39. More cattle. Based on the model $N(1152, 84)$ describing Angus steer weights, what are the cutoff values for

a) the highest 10% of the weights?
b) the lowest 20% of the weights?
c) the middle 40% of the weights?

40. More IQs. In the Normal model $N(100, 16)$, what cutoff value bounds

a) the highest 5% of all IQs?
b) the lowest 30% of the IQs?
c) the middle 80% of the IQs?

41. Cattle, finis. Consider the Angus weights model *N*(1152, 84) one last time.

a) What weight represents the 40th percentile?

b) What weight represents the 99th percentile?

c) What's the IQR of the weights of these Angus steers?

42. IQ, finis. Consider the IQ model *N*(100, 16) one last time.

a) What IQ represents the 15th percentile?

b) What IQ represents the 98th percentile?

c) What's the IQR of the IQs?

43. Cholesterol. Assume the cholesterol levels of adult American women can be described by a Normal model with a mean of 188 mg/dL and a standard deviation of 24.

a) Draw and label the Normal model.

b) What percent of adult women do you expect to have cholesterol levels over 200 mg/dL?

c) What percent of adult women do you expect to have cholesterol levels between 150 and 170 mg/dL?

d) Estimate the IQR of the cholesterol levels.

e) Above what value are the highest 15% of women's cholesterol levels?

44. Tires. A tire manufacturer believes that the treadlife of its snow tires can be described by a Normal model with a mean of 32,000 miles and standard deviation of 2500 miles.

a) If you buy a set of these tires, would it be reasonable for you to hope they'll last 40,000 miles? Explain.

b) Approximately what fraction of these tires can be expected to last less than 30,000 miles?

c) Approximately what fraction of these tires can be expected to last between 30,000 and 35,000 miles?

d) Estimate the IQR of the treadlives.

e) In planning a marketing strategy, a local tire dealer wants to offer a refund to any customer whose tires fail to last a certain number of miles. However, the dealer does not want to take too big a risk. If the dealer is willing to give refunds to no more than 1 of every 25 customers, for what mileage can he guarantee these tires to last?

45. Kindergarten. Companies that design furniture for elementary school classrooms produce a variety of sizes for kids of different ages. Suppose the heights of kindergarten children can be described by a Normal model with a mean of 38.2 inches and standard deviation of 1.8 inches.

a) What fraction of kindergarten kids should the company expect to be less than 3 feet tall?

b) In what height interval should the company expect to find the middle 80% of kindergarteners?

c) At least how tall are the biggest 10% of kindergarteners?

46. Body temperatures. Most people think that the "normal" adult body temperature is 98.6°F. That figure, based on a 19th-century study, has recently been challenged. In a 1992 article in the *Journal of the American Medical Association,* researchers reported that a more accurate figure may be 98.2°F. Furthermore, the standard deviation appeared to be around 0.7°F. Assume that a Normal model is appropriate.

a) In what interval would you expect most people's body temperatures to be? Explain.

b) What fraction of people would be expected to have body temperatures above 98.6°F?

c) Below what body temperature are the coolest 20% of all people?

47. Eggs. Hens usually begin laying eggs when they are about 6 months old. Young hens tend to lay smaller eggs, often weighing less than the desired minimum weight of 54 grams.

a) The average weight of the eggs produced by the young hens is 50.9 grams, and only 28% of their eggs exceed the desired minimum weight. If a Normal model is appropriate, what would the standard deviation of the egg weights be?

b) By the time these hens have reached the age of 1 year, the eggs they produce average 67.1 grams, and 98% of them are above the minimum weight. What is the standard deviation for the appropriate Normal model for these older hens?

c) Are egg sizes more consistent for the younger hens or the older ones? Explain.

48. Tomatoes. Agricultural scientists are working on developing an improved variety of Roma tomatoes. Marketing research indicates that customers are likely to bypass Romas that weigh less than 70 grams. The current variety of Roma plants produces fruit that averages 74 grams, but 11% of the tomatoes are too small. It is reasonable to assume that a Normal model applies.

a) What is the standard deviation of the weights of Romas now being grown?

b) Scientists hope to reduce the frequency of undersized tomatoes to no more than 4%. One way to accomplish this is to raise the average size of the fruit. If the standard deviation remains the same, what target mean should they have as a goal?

c) The researchers produce a new variety with a mean weight of 75 grams, which meets the 4% goal. What is the standard deviation of the weights of these new Romas?

d) Based on their standard deviations, compare the tomatoes produced by the two varieties.

JUST CHECKING Answers

1. a) On the first test, the mean is 88 and the SD is 4, so $z = (90 - 88)/4 = 0.5$. On the second test, the mean is 75 and the SD is 5, so $z = (80 - 75)/5 = 1.0$. The first test has the lower z-score, so it is the one that will be dropped.

b) No. The second test is 1 standard deviation above the mean, farther away than the first test, so it's the better score relative to the class.

2. a) The mean would increase to 500.

b) The standard deviation is still 100 points.

c) The two boxplots would look nearly identical (the shape of the distribution would remain the same), but the later one would be shifted 50 points higher.

3. The standard deviation is now 2.54 millimeters, which is the same as 0.1 inches. Nothing has changed. The standard deviation has "increased" only because we're reporting it in millimeters now, not inches.

4. The mean is 184 centimeters, with a standard deviation of 8 centimeters. 2 meters is 200 centimeters, which is 2 standard deviations above the mean. We expect 5% of the men to be more than 2 standard deviations below or above the mean, so half of those, 2.5%, are likely to be above 2 meters.

5. a) We know that 68% of the time we'll be within 1 standard deviation (2 min) of 20. So 32% of the time we'll arrive in less than 18 or more than 22 minutes. Half of those times (16%) will be greater than 22 minutes, so 84% will be less than 22 minutes.

b) 24 minutes is 2 standard deviations above the mean. Because of the 95% rule, we know 2.5% of the times will be more than 24 minutes.

c) Traffic incidents may occasionally increase the time it takes to get to school, so the driving times may be skewed to the right, and there may be outliers.

d) If so, the Normal model would not be appropriate and the percentages we predict would not be accurate.

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

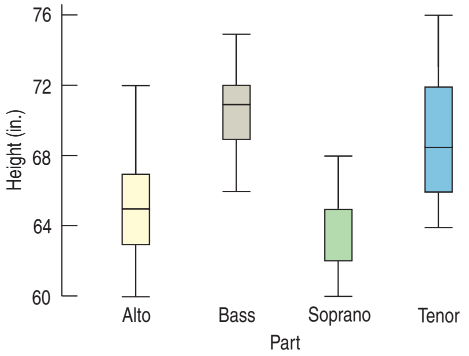
a) Display these data with an appropriate graph.
b) Report appropriate summary statistics.
c) Write a few sentences about this distribution.

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.

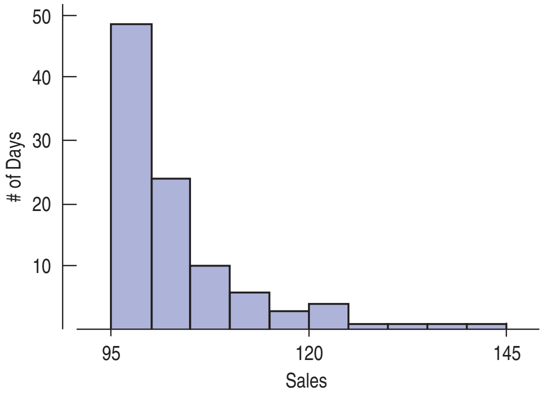


a) It appears that the median height for sopranos is missing, but actually the median and the upper quartile are equal. How could that happen?
b) Write a few sentences describing what you see.

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.

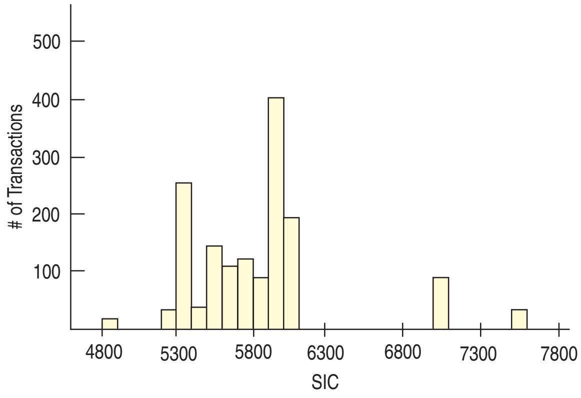
7. State University. Public relations staff at State U. collected data on people's opinions of various colleges and universities in their state. They phoned 850 local residents. After identifying themselves, the callers asked the survey participants their ages, whether they had attended college, and 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?

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.

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.

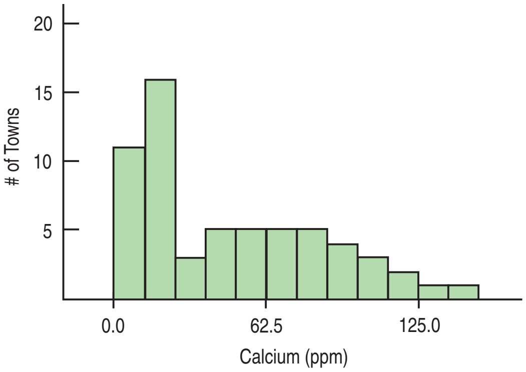
[2] Since 1997 the SIC has been replaced by the NAICS, a code of six letters.

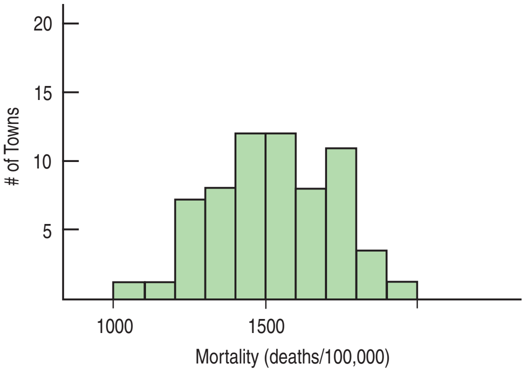
Miles from Home	% of Accidents
Less than 1	23
1 to 5	29
6 to 10	17
11 to 15	8
16 to 20	6
Over 20	17

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

T **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.

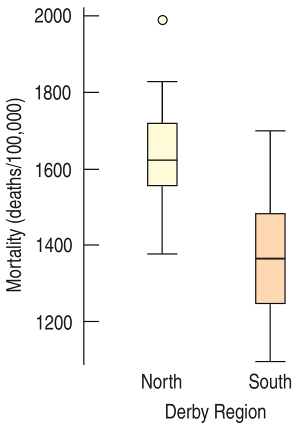




T **16. Hard water II.** The data set from England and Wales also notes for each town whether it was south or north of Derby. Here are some summary statistics and a comparative boxplot for the two regions.

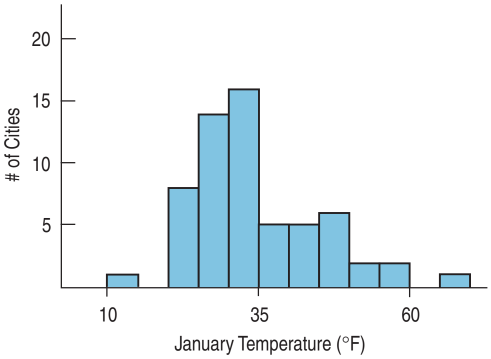
Summary of Mortality

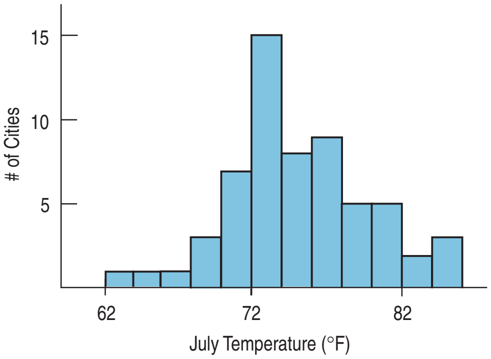
Group	Count	Mean	Median	StdDev
North	34	1631.59	1631	138.470
South	27	1388.85	1369	151.114



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

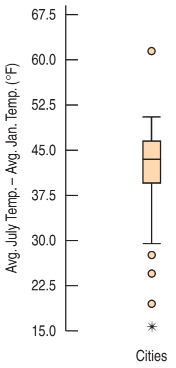
17. Seasons. Average daily temperatures in January and July for 60 large U.S. cities are graphed in the histograms below.





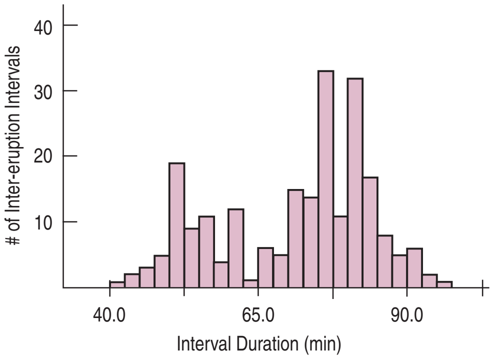
a) What aspect of these histograms makes it difficult to compare the distributions?

b) What differences do you see between the distributions of January and July average temperatures?



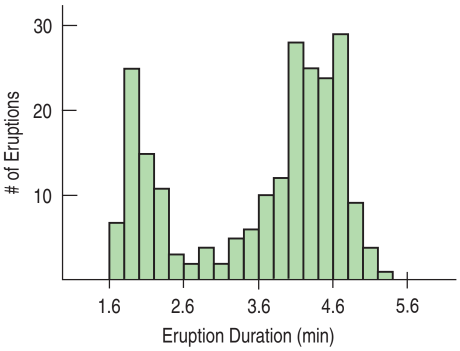
c) Differences in temperatures (July–January) for each of the cities are displayed in the boxplot above. Write a few sentences describing what you see.

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.



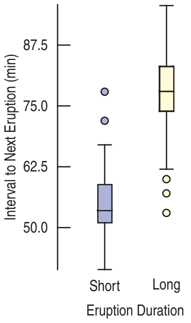
19. Old Faithful? Does the duration of an eruption have an effect on the length of time that elapses before the next eruption?

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.

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.

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.

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.

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*				
		1	2	3	4+	Total
Major	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.

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.

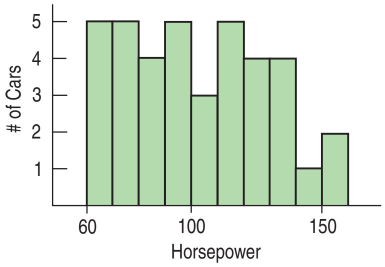
31. Engines. One measure of the size of an automobile engine is its "displacement," the total volume (in liters or cubic inches) of its cylinders. Summary statistics for several models of new cars are shown. These displacements were measured in cubic inches.

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.

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	1234
0	111123444
0	5555679
1	00113
1	
2	2
2	5

Profits (% of sales)

(–0|3 means a loss of 3%)

a) Find the 5-number summary.

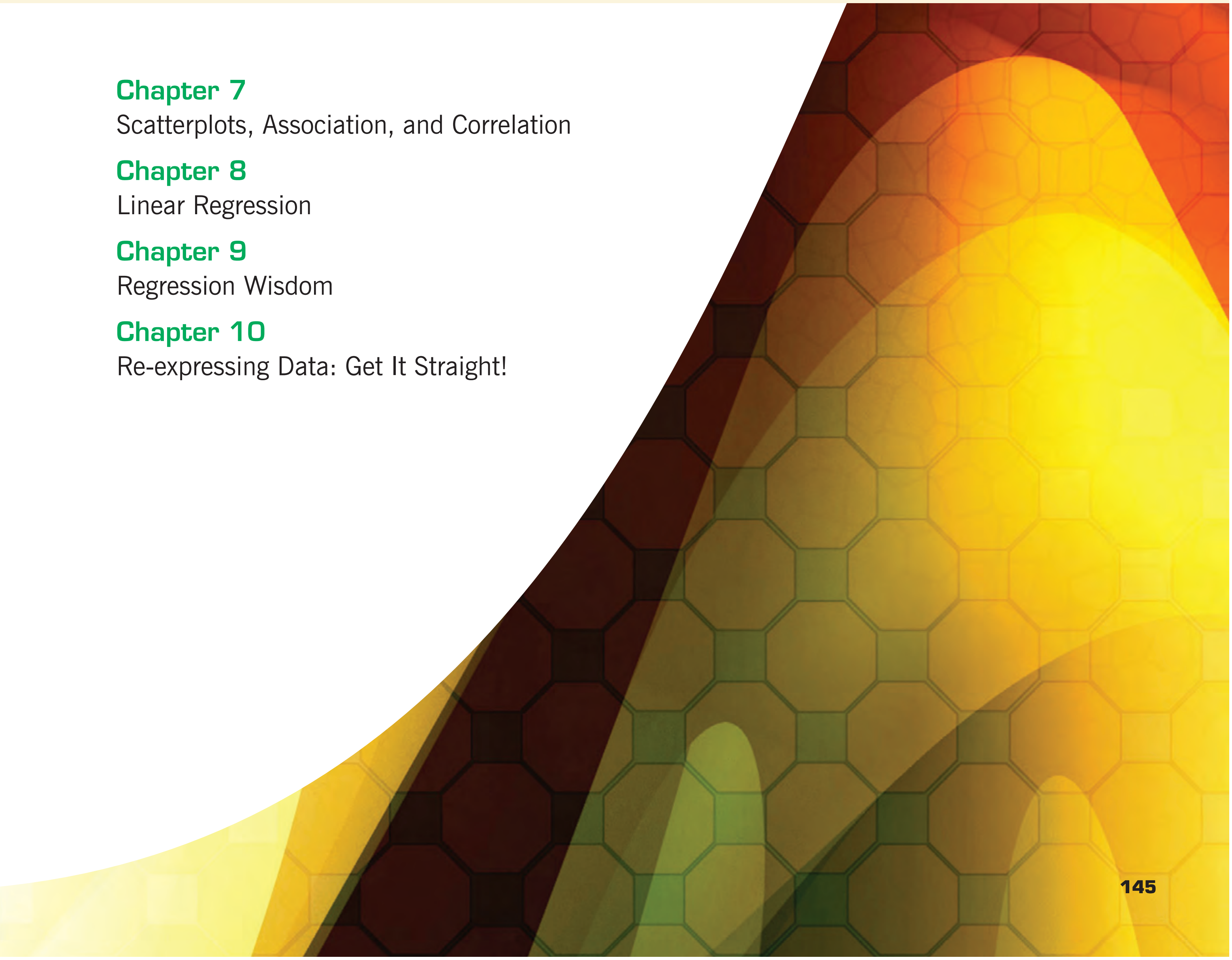
b) Draw a boxplot for these data.

c) Find the mean and standard deviation.

d) Describe the distribution of profits for these corporations.

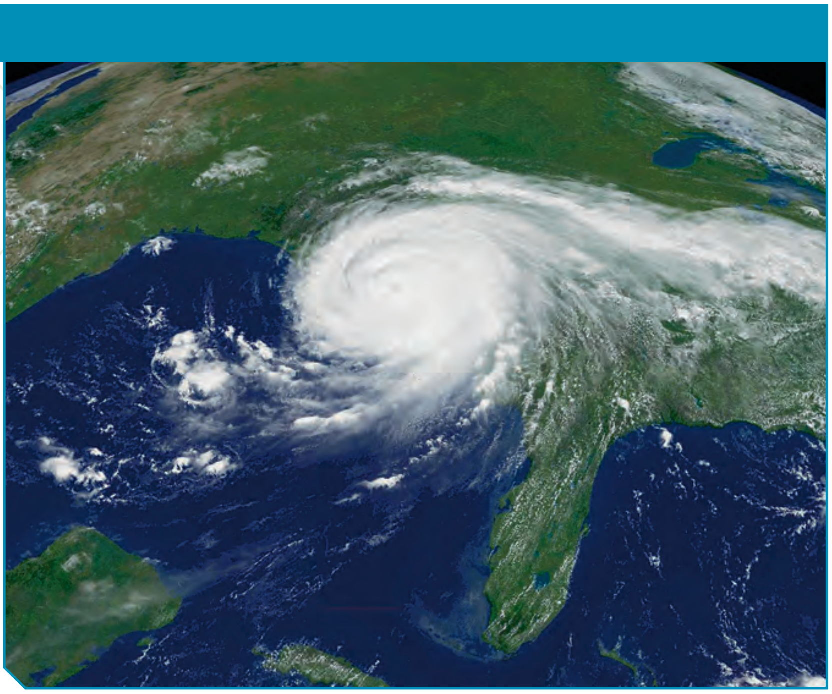
PART II

Exploring Relationships Between Variables



CHAPTER 7

Scatterplots, Association, and Correlation



WHO	Years 1970–2005
WHAT	Mean error in the position of Atlantic hurricanes as predicted 72 hours ahead by the NHC
UNITS	nautical miles
WHEN	1970–2005
WHERE	Atlantic and Gulf of Mexico
WHY	The NHC wants to improve prediction models

Hurricane Katrina killed 1,836 people[1] and caused well over 100 billion dollars in damage—the most ever recorded. Much of the damage caused by Katrina was due to its almost perfectly deadly aim at New Orleans.

Where will a hurricane go? People want to know if a hurricane is coming their way, and the National Hurricane Center (NHC) of the National Oceanic and Atmospheric Administration (NOAA) tries to predict the path a hurricane will take. But hurricanes tend to wander around aimlessly and are pushed by fronts and other weather phenomena in their area, so they are notoriously difficult to predict. Even relatively small changes in a hurricane's track can make big differences in the damage it causes.

To improve hurricane prediction, NOAA[2] relies on sophisticated computer models, and has been working for decades to improve them. How well are they doing? Have predictions improved in recent years? Has the improvement been consistent? Here's a timeplot of the mean error, in nautical miles, of the NHC's 72-hour predictions of Atlantic hurricanes since 1970:

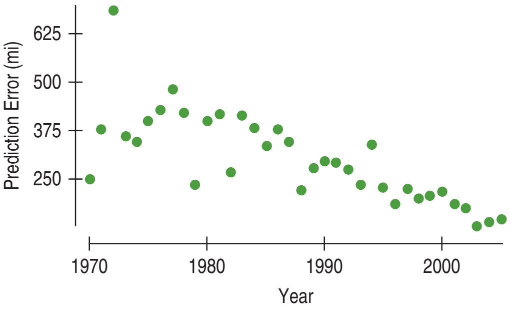


FIGURE 7.1

A scatterplot of the average error in nautical miles of the predicted position of Atlantic hurricanes for predictions made by the National Hurricane Center of NOAA, plotted against the Year in which the predictions were made.

Look, Ma, no origin!
Scatterplots usually don't—and shouldn't—show the origin, because often neither variable has values near 0. The display should focus on the part of the coordinate plane that actually contains the data. In our example about hurricanes, none of the prediction errors or years were anywhere near 0, so the computer drew the scatterplot with axes that don't quite meet.

[1] In addition, 705 are still listed as missing.
[2] www.nhc.noaa.gov