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.

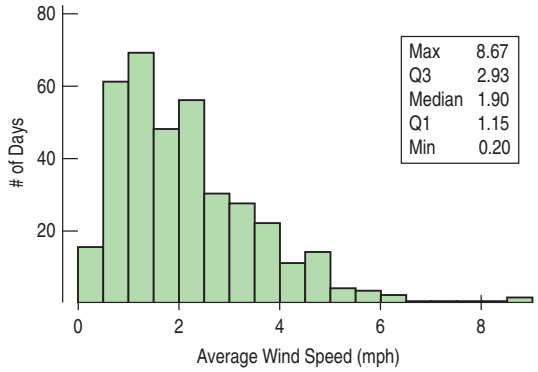
One of the variables measured in the forest is wind speed. Three remote anemometers generate far too much data to report, so, as summaries, you'll find the minimum, maximum, and average wind speed (in mph) for each day.

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



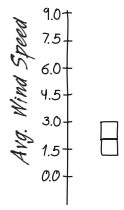
Let's start with the "big picture." Here's a histogram and 5-number summary of the *Average Wind Speed* for every day in 1989. Because of the skewness, we'll report the median and IQR. We can see that the distribution of *Average Wind Speed* is unimodal and skewed to the right. Median daily wind speed 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

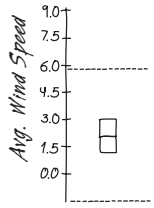
A histogram of daily Average Wind Speed *for 1989. It is unimodal and skewed to the right, with a possible high outlier.*

Boxplots and 5-Number Summaries

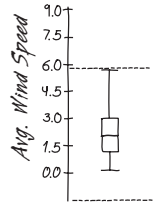
1.



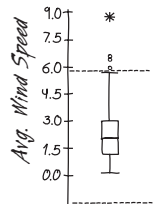
2.



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.

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 answerd 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.

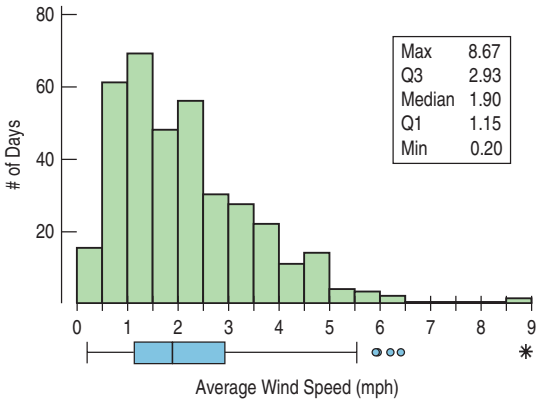


FIGURE 5.2

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):

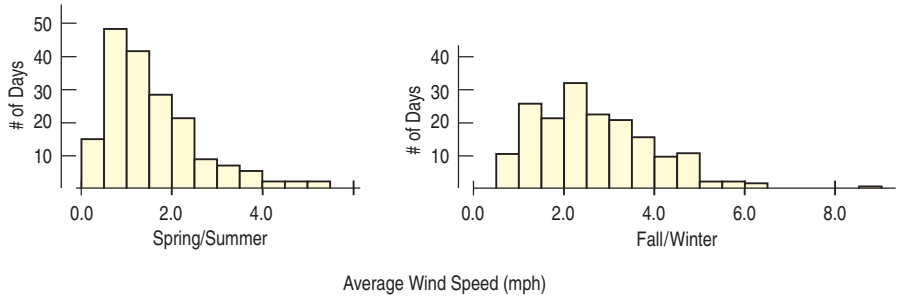


FIGURE 5.3

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:

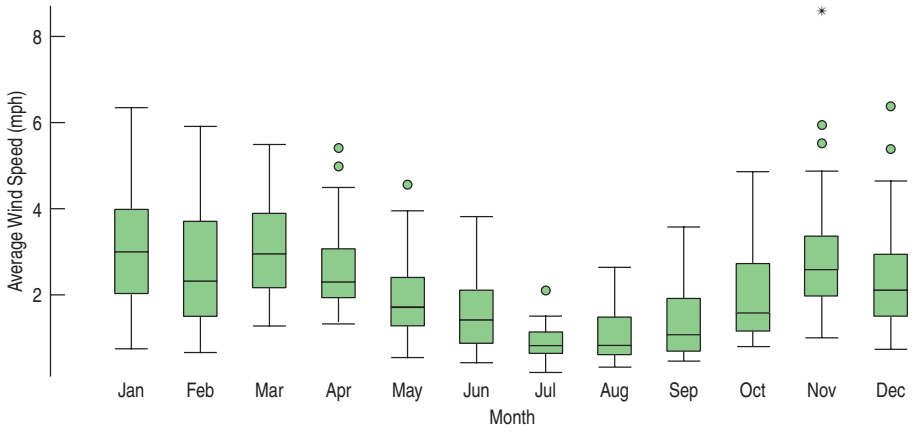


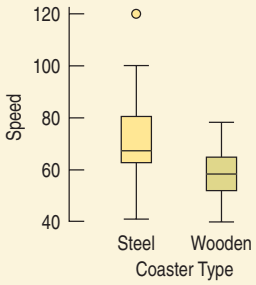
FIGURE 5.4
Boxplots of the average daily wind speed for each month show seasonal patterns in both the centers and spreads.

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.

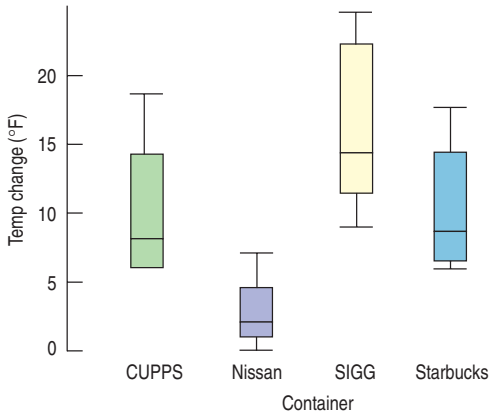
✓ **Quantitative Data Condition:** The *Temperature Changes* are quantitative, with units of °F. Boxplots are appropriate displays for comparing the groups. Numerical summaries of each group are appropriate as well.

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.



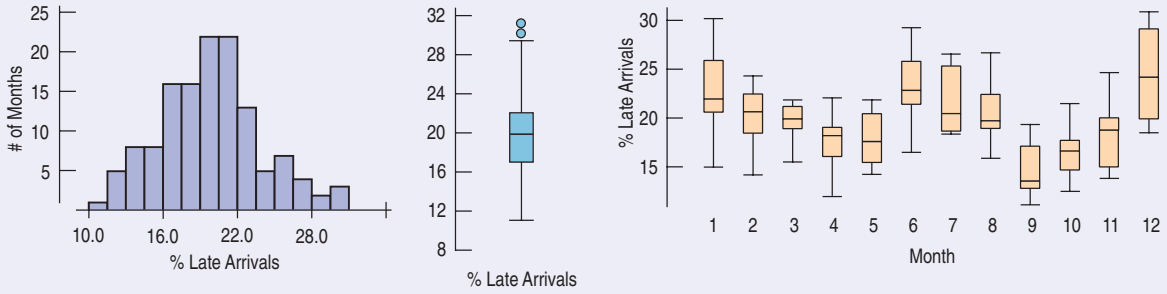
TELL

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:

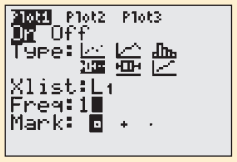
Boys: 22, 17, 18, 29, 22, 22, 23, 24, 23, 17, 21

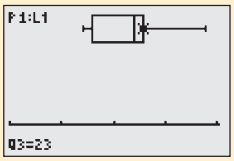
Girls: 25, 20, 12, 19, 28, 24, 22, 21, 25, 26, 25, 16, 27, 22

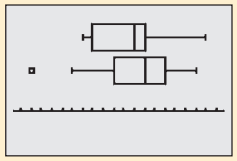
Enter these data in `L1` (*Boys*) and `L2` (*Girls*).

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

- 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!

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.

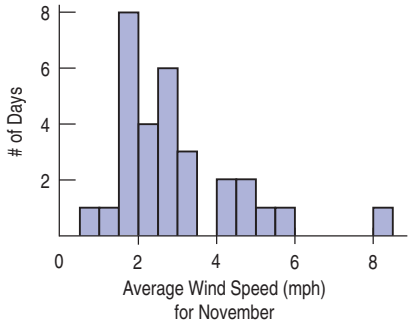


FIGURE 5.5

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.

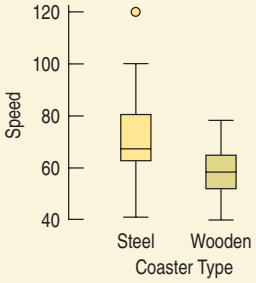
FOR EXAMPLE **Checking out the outliers**

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:

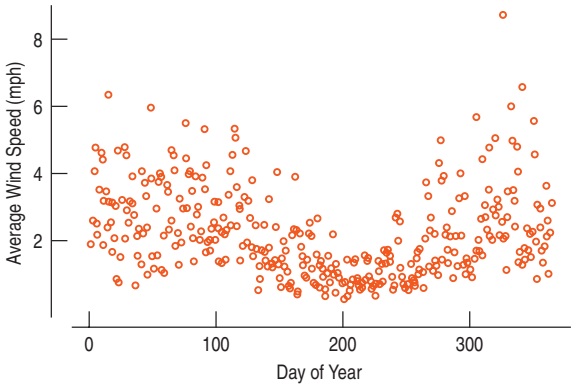


FIGURE 5.6
A timeplot of Average Wind Speed *shows the overall pattern and changes in variation.*

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:

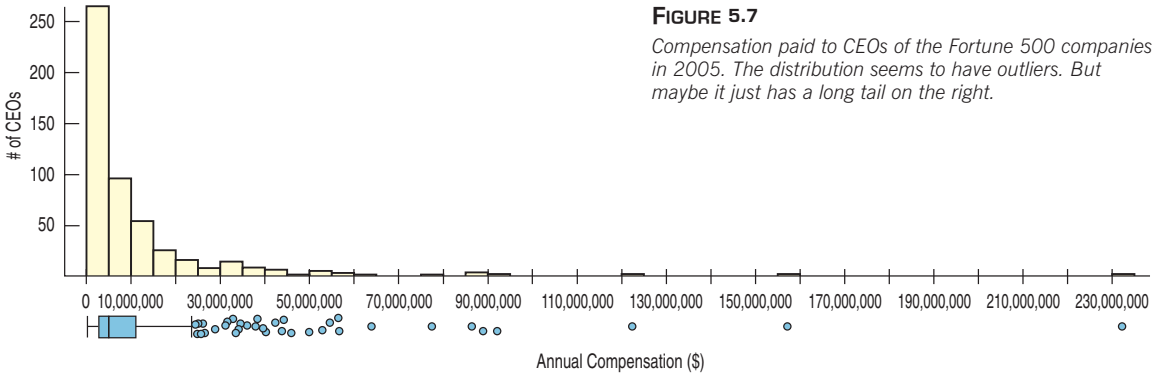


FIGURE 5.7
Compensation paid to CEOs of the Fortune 500 companies in 2005. The distribution seems to have outliers. But maybe it just has a long tail on the right.

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:

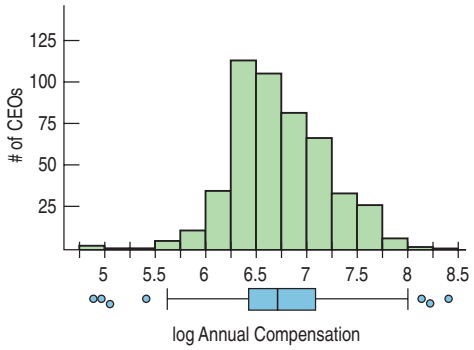


FIGURE 5.8
The logarithms of 2005 CEO compensations are much more nearly symmetric.

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.

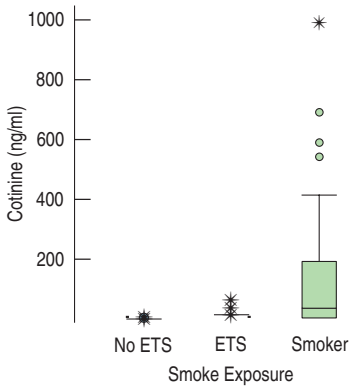


FIGURE 5.9

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?

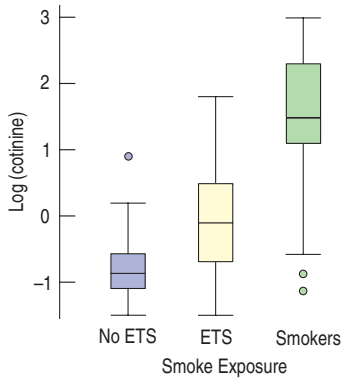


FIGURE 5.10

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.

  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

- 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.

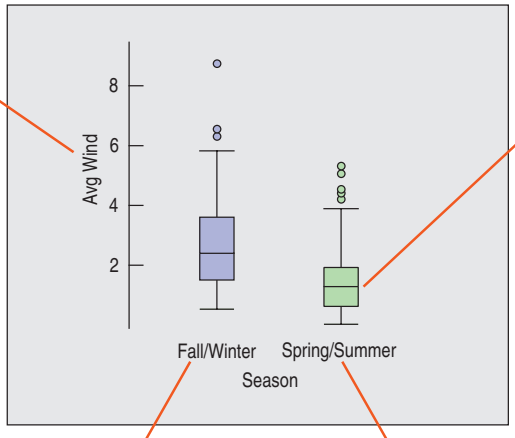
TELL

- 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

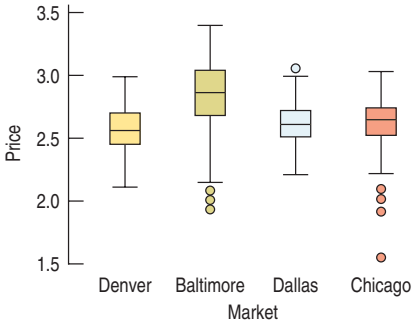
1. **In the news.** Find an article in a newspaper, magazine, or the Internet that compares two or more groups of data.
   a) Does the article discuss the W's?
   b) Is the chosen display appropriate? Explain.
   c) Discuss what the display reveals about the groups.
   d) Does the article accurately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

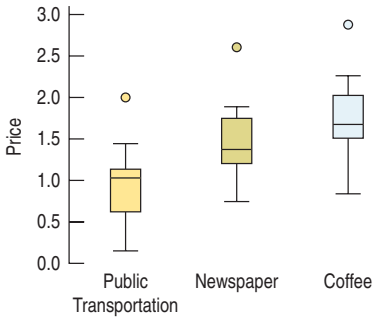
4. **Groups on the Internet.** Find data on the Internet (or elsewhere) for two or more groups. Make appropriate displays to compare the groups, and interpret what you find.

T 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:



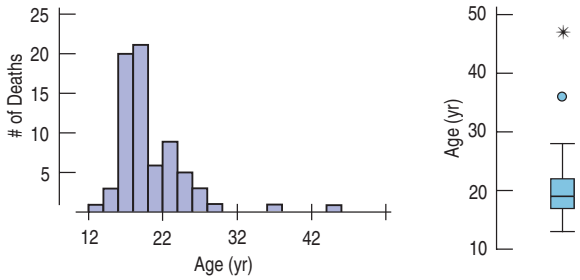
   a) Do prices appear to be the same in the four markets? Explain.
   b) Does the presence of any outliers affect your overall conclusions about prices in the four markets?

T 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).



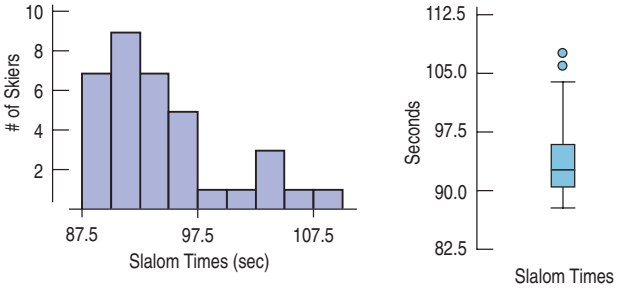
   a) On average, which commodity is the most expensive?
   b) Is a newspaper always more expensive than a ride on public transportation? Explain.
   c) Does the presence of outliers affect your conclusions in a) or b)?

T 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:



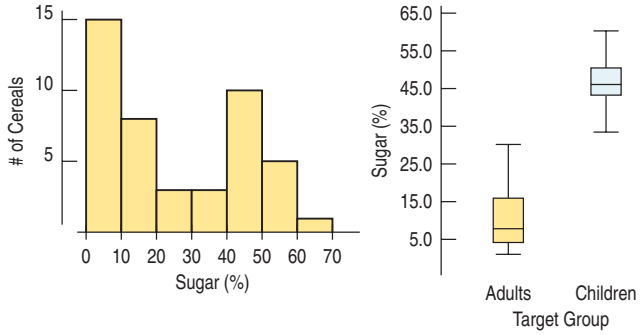
   a) What features of the distribution can you see in both the histogram and the boxplot?
   b) What features of the distribution can you see in the histogram that you could not see in the boxplot?
   c) What summary statistic would you choose to summarize the center of this distribution? Why?
   d) What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

T 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:



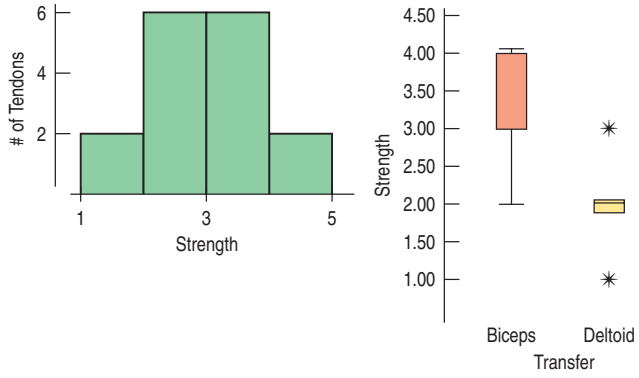
a) What features of the distribution can you see in both the histogram and the boxplot?
b) What features of the distribution can you see in the histogram that you could not see in the boxplot?
c) What summary statistic would you choose to summarize the center of this distribution? Why?
d) What summary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

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.

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.

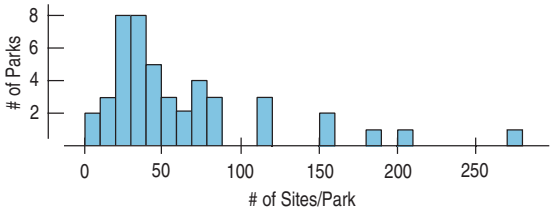
11. Population growth. Here is a "back-to-back" stem-and-leaf display that shows two data sets at once—one going to the left, one to the right. The display compares the percent change in population for two regions of the United States (based on census figures for 1990 and 2000). The fastest growing states 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.

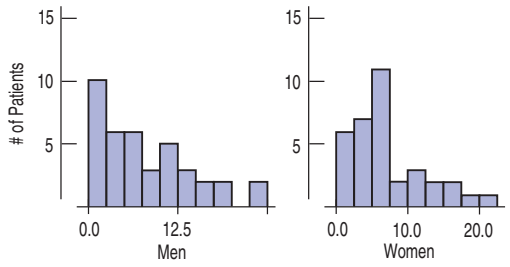
12. Camp sites. Shown below are the histogram and summary statistics for the number of camp sites at public parks in Vermont.



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

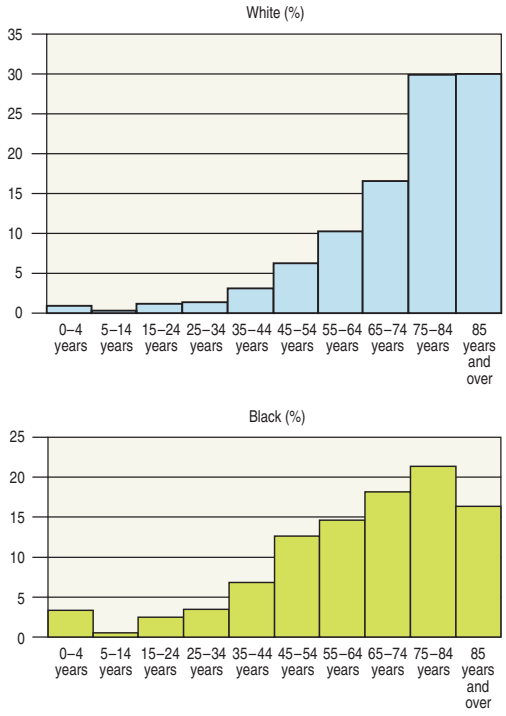
a) Which statistics would you use to identify the center and spread of this distribution? Why?
b) How many parks would you classify as outliers? Explain.
c) Create a boxplot for these data.
d) Write a few sentences describing the distribution.

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.



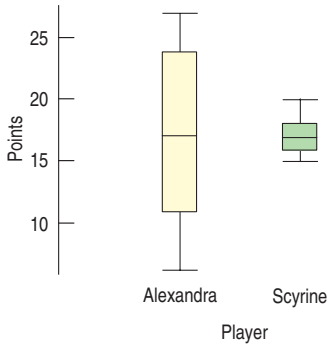
a) What would you suggest be changed about these histograms to make them easier to compare?
b) Describe these distributions by writing a few sentences comparing the duration of hospitalization for men and women.
c) Can you suggest a reason for the peak in women's length of stay?

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:



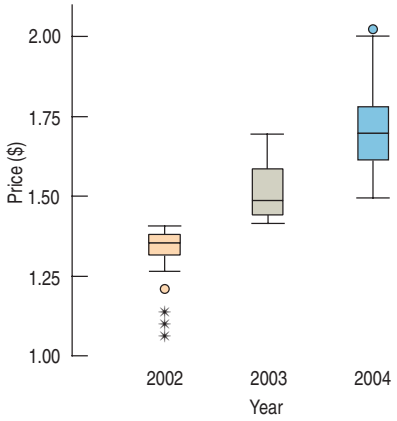
a) Describe the overall shapes of these distributions.
b) How do the distributions differ?
c) Look carefully at the bar definitions. Where do these plots violate the rules for statistical graphs?

T **15. Women's basketball.** Here are boxplots of the points scored during the first 10 games of the season for both Scyrine and Alexandra:



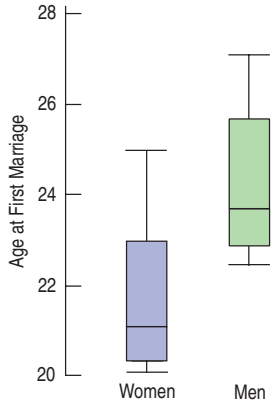
a) Summarize the similarities and differences in their performance so far.
b) The coach can take only one player to the state championship. Which one should she take? Why?

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

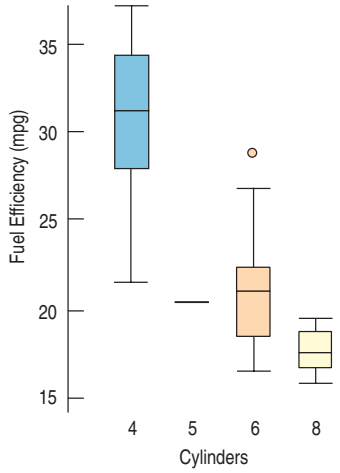


a) Compare the distribution of prices over the three years.
b) In which year were the prices least stable? Explain.

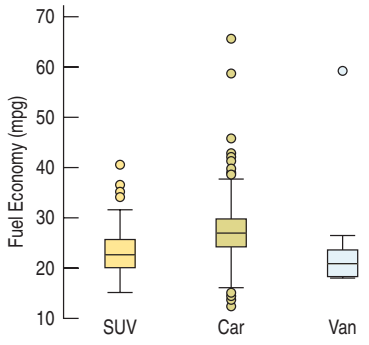
T **17. Marriage age.** In 1975, did men and women marry at the same age? Here are boxplots of the age at first marriage for a sample of U.S. citizens then. Write a brief report discussing what these data show.



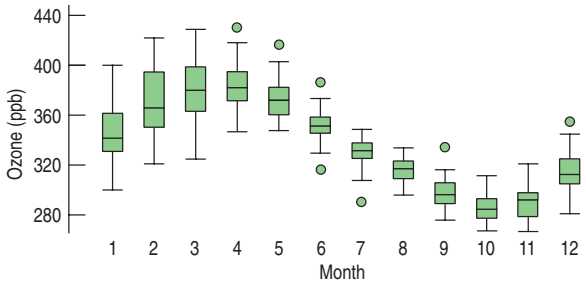
18. Fuel economy. Describe what these boxplots tell you about the relationship between the number of cylinders a car's engine has and the car's fuel economy (mpg):



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.

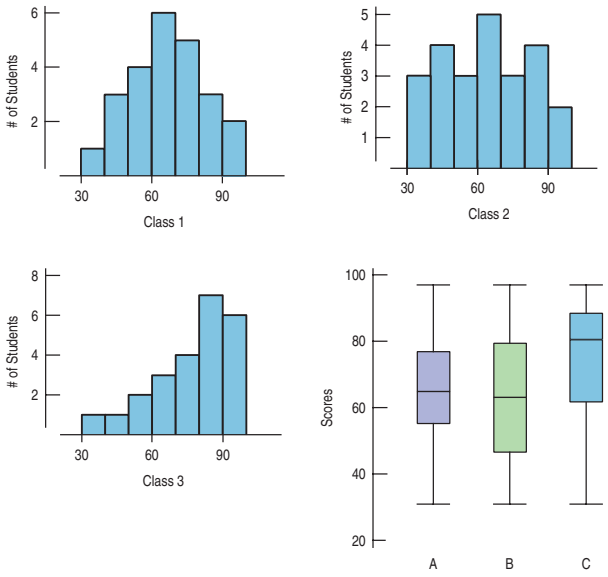


20. Ozone. Ozone levels (in parts per billion, ppb) were recorded at sites in New Jersey monthly between 1926 and 1971. Here are boxplots of the data for each month (over the 46 years), lined up in order (January = 1):



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.

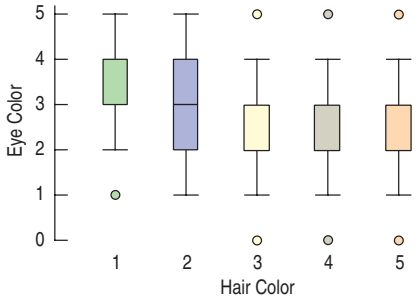
21. Test scores. Three Statistics classes all took the same test. Histograms and boxplots of the scores for each class are shown below. Match each class with the corresponding boxplot.



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

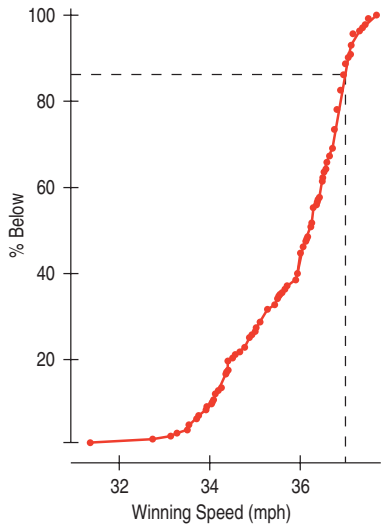
a) Describe the W's for these data.
b) Name the variables and classify each as categorical or quantitative.
c) Create parallel boxplots to display these results as best you can with this information.
d) Write a few sentences comparing the performances of the three groups.

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

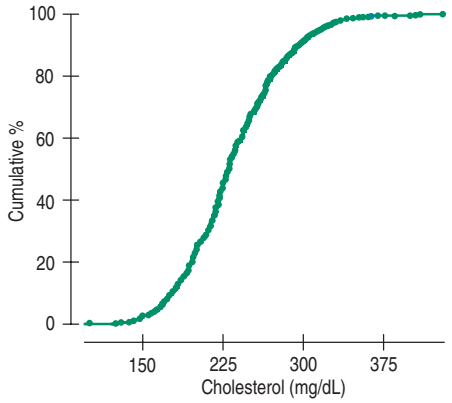
a) Create parallel boxplots comparing the scores of boys and girls as best you can from the information given.
b) Write a brief report on these results. Be sure to discuss the shape, center, and spread of the scores.

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.

T **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?

T **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.

T **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.

T **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

T **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

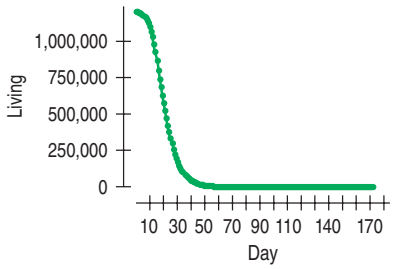
a) Create graphical displays for these two groups.
b) Write a few sentences comparing the distributions.

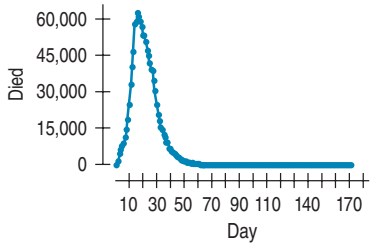
T **34. Baseball.** American League baseball teams play their games with the designated hitter rule, meaning that pitchers do not bat. The League believes that replacing the pitcher, typically a weak hitter, with another player in the batting order produces more runs and generates more interest among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

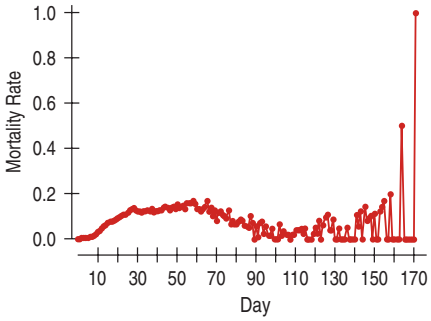
Average Runs	League	Average Runs	League
11.1	American	14.0	National
10.8	American	11.6	National
10.8	American	10.4	National
10.3	American	10.9	National
10.3	American	10.2	National
10.1	American	9.5	National
10.0	American	9.5	National
9.5	American	9.5	National
9.4	American	9.5	National
9.3	American	9.1	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.

T **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.







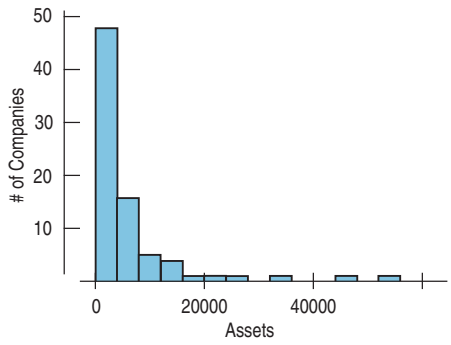
a) On approximately what day did the most flies die?
b) On what day during the first 100 days did the largest *proportion* of flies die?
c) When did the number of fruit flies alive stop changing very much from day to day?

T **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

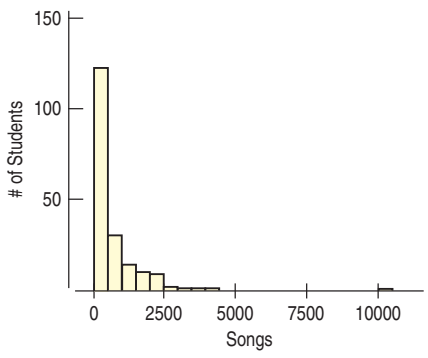
a) Create a stem-and-leaf display or a histogram of these data.
b) Create a timeplot.
c) Using features apparent in the stem-and-leaf display (or histogram) and the timeplot, write a few sentences about deaths caused by drunk driving.

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:



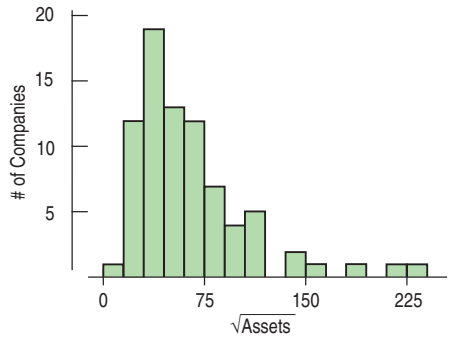
a) What aspect of this distribution makes it difficult to summarize, or to discuss, center and spread?
b) What would you suggest doing with these data if we want to understand them better?

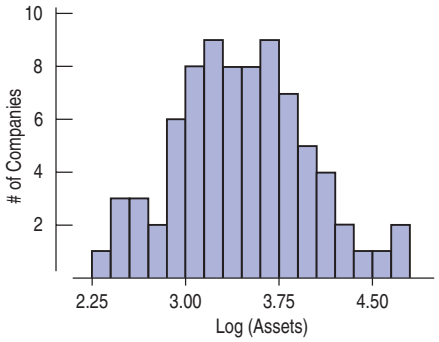
38. Music library. Students were asked how many songs they had in their digital music libraries. Here's a display of the responses:



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:

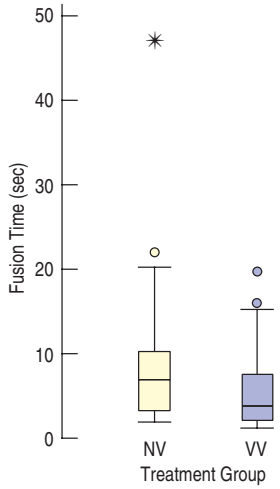
2745	703	302	242	119	40	7
1697	489	274	200	118	32	4
1656	430	274	198	115	31	
978	334	255	129	92	17	

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.

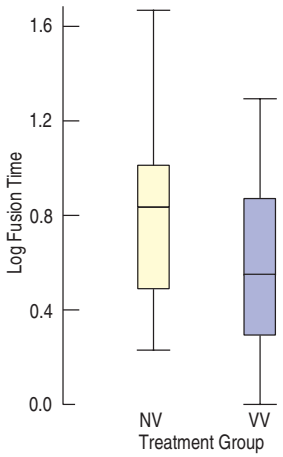
T **41. Stereograms.** Stereograms appear to be composed entirely of random dots. However, they contain separate images that a viewer can "fuse" into a three-dimensional (3D) image by staring at the dots while defocusing the eyes. An experiment was performed to determine whether knowledge of the embedded 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.