Displaying and Describing Quantitative Data

CHAPTER 5

AIG

The American International Group (AIG) was once the 18th largest corporation in the world. AIG was founded nearly 100 years ago by Cornelius Vander Starr who opened an insurance agency in Shanghai, China. As the first Westerner to sell insurance to the Chinese, Starr grew his business rapidly until 1949 when Mao Zedong and the People's Liberation Army took over Shanghai. Starr moved the company to New York City, where it continued to grow, expanding its markets worldwide. In 2004, AIG stock hit an all-time high of $76.77, putting its market value at nearly $300 billion.

According to its own website, "By early 2007 AIG had assets of $1 trillion, $110 billion in revenues, 74 million customers and 116,000 employees in 130 countries and jurisdictions. Yet just 18 months later, AIG found itself on the brink of failure and in need of emergency government assistance." AIG was one of the largest beneficiaries of the U.S. government's Troubled Asset Relief Program (TARP), established in 2008 during the financial crisis to purchase assets and equity from financial institutions. TARP was an attempt to strengthen the financial sector and avoid a repeat of a depression as severe as the 1930s. Many banks quickly repaid the government part or all of the money given to them under the TARP program, but AIG, which received $170 billion, did not.

By 2009, AIG stock had lost more than 99% of its value, hitting $0.35 in early March. That same month AIG became embroiled in controversy when it disclosed that it had paid $218 million in bonuses to employees of its financial services division. AIG's drop in stock price represented a loss of nearly $300 billion for investors. Portfolio managers typically examine stock prices and volumes to determine stock volatility and to help them decide which stocks to buy and sell. Were there early warning signs in AIG's data?

To learn more about the behavior and volatility of AIG's stock, let's start by looking at Table 5.1, which gives the monthly average stock price (in dollars) for the six years leading up to the company's crisis.

	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
2002	77.26	72.95	73.72	71.57	68.42	65.99	61.22	64.10	58.04	60.26	65.03	59.96
2003	59.74	49.57	49.41	54.38	56.52	57.88	59.80	61.51	59.39	60.93	58.73	62.37
2004	69.02	73.25	72.06	74.21	70.93	72.61	69.85	69.58	70.67	62.31	62.17	65.33
2005	66.74	68.96	61.55	51.77	53.81	55.66	60.27	60.86	60.54	62.64	67.06	66.72
2006	68.33	67.02	67.15	64.29	63.14	59.74	59.40	62.00	65.25	67.02	69.86	71.35
2007	70.45	68.99	68.14	68.25	71.78	71.75	68.64	65.21	66.02	66.12	56.86	58.13

Table 5.1 Monthly stock price in dollars of AIG stock for the period 2002 through 2007.

It's hard to tell very much from tables of values like this. You might get a rough idea of how much the stock cost—usually somewhere around $60 or so, but that's about it.

5.1 Displaying Quantitative Variables

Instead, let's follow the first rule of data analysis and make a picture. What kind of picture should we make? It can't be a bar chart or a pie chart. Those are only for categorical variables, and AIG's stock price is a *quantitative* variable, whose units are dollars.

WHO	Months
WHAT	Monthly average price for AIG's stock (in dollars)
WHEN	2002 through 2007
WHERE	New York Stock Exchange
WHY	To examine AIG stock volatility

Histograms

Here are the monthly prices of AIG stock displayed in a histogram.

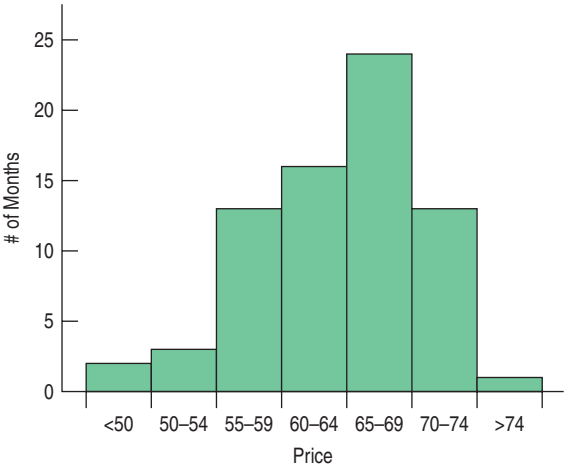


Figure 5.1 Monthly average prices of AIG stock. The histogram displays the distribution of prices by showing for each "bin" of prices, the number of months having prices in that bin.

Like a bar chart, a **histogram** plots the bin counts as the heights of bars. It counts the number of cases that fall into each bin, and displays that count as the height of the corresponding bar. In this histogram of monthly average prices, each bin has a width of $5, so, for example, the height of the tallest bar says that there were 24 months whose average price of AIG stock was between $65 and $70. In this way, the histogram displays the entire distribution of prices at a glance. Unlike a bar chart, which puts gaps between bars to separate the categories, there are no gaps between the bars of a histogram unless there are actual gaps in the data. **Gaps** indicate a region where there are no values. Gaps can be important features of the distribution so watch out for them and point them out.

For categorical variables, each category got its own bar. The only choice was whether to combine categories for ease of display. For quantitative variables, we have to choose the width of the bins.

- **Making a Histogram by Hand** Although you'll rarely make a histogram by hand, it can be instructive to see how it might be done. Some of the same choices that you have to make by hand will either be made by software automatically, or with input from you.

  Step 1. Organize your data into a table. Divide the data into equal intervals or bins so that all values are covered. The number of bins depends on how many data values you have. For small (fewer than 25 or so values) data sets, 5 bins is fine. For large data sets you may need 20 or more.

  You will probably want the width of intervals to be aesthetically pleasing (bins that have widths that end with 5's or 0's are popular—for example, 35–40,40–45, 45–50, etc). Now create two columns—one for the bins and the other for the frequencies. This creates a frequency distribution much like that for a categorical variable, but here instead of categories, we have bins of equal width. You'll have to decide whether to put values that lie at the end points of the bin into the left or right bin. Most software histogram programs put values into the bin on the right with the larger values, so 40 would go into the bin 40–45, not in the bin 35–40, but

Price Bin	# of Months
45–50	2
50–55	3
55–60	13
60–65	16
65–70	24
70–75	13
75–80	1

either choice is possible. In fact, Excel chooses to put the values to the left, so that 5 goes into the bin 0–5 not the bin 5–10.

Step 2. On a sheet of paper, mark the bins on the *x*-axis (horizontal axis) with no spaces between the bins. Mark the frequencies on the *y*-axis (vertical axis) over the center of each bin.

Step 3. Plot your data. For each bin, draw a horizontal line at the appropriate frequency over the bin. Then, draw vertical bars on the sides of each bin, reaching up to the corresponding frequency.

From the histogram, we can see that in these months a typical AIG stock price was near $60 or so. We can see that although they vary, most of the monthly prices were between $55 and $75. Only in a very few months was the average price below $55. It's important to note that the histogram is a static picture. We have treated these prices simply as a collection of months, with no sense of time, and shown their distribution. Later in the chapter we will add time to the story.

Does the distribution look as you expected? It's often a good idea to imagine what the distribution might look like before making the display. That way you're less likely to be fooled by errors either in your display or in the data themselves.

If our focus is on the overall pattern of how the values are distributed rather than on the counts themselves, it can be useful to make a relative frequency histogram, replacing the counts on the vertical axis with the percentage of the total number of cases falling in each bin (simply divide the counts in each bin by the total number of data values). The shape of the histogram is exactly the same; only the labels are different. A **relative frequency histogram** is faithful to the area principle by displaying the *percentage* of cases in each bin instead of the count.

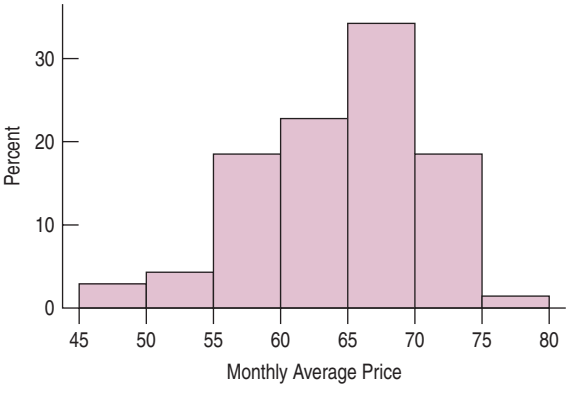


Figure 5.2 A relative frequency histogram looks just like a frequency histogram except that the *y*-axis now shows the percentage of months in each bin.

For Example — Creating a histogram

1. As the chief financial officer of a music download site, you've just secured the rights to offer downloads of a new album. You'd like to see how well it's selling, so you collect the number of downloads per hour for the past 24 hours:

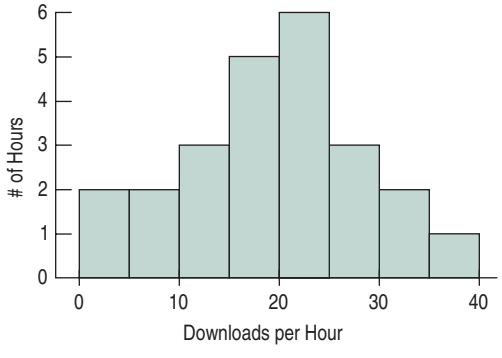
Hour	Downloads	Hour	Downloads
12:00 a.m.	36	12:00 p.m.	25
1:00 a.m.	28	1:00 p.m.	22
2:00 a.m.	19	2:00 p.m.	17
3:00 a.m.	10	3:00 p.m.	18
4:00 a.m.	5	4:00 p.m.	20
5:00 a.m.	3	5:00 p.m.	23
6:00 a.m.	2	6:00 p.m.	21
7:00 a.m.	6	7:00 p.m.	18
8:00 a.m.	12	8:00 p.m.	24
9:00 a.m.	14	9:00 p.m.	30
10:00 a.m.	20	10:00 p.m.	27
11:00 a.m.	18	11:00 p.m.	30

Question: Make a histogram for this variable.

Answer: Create a frequency table of bins of width five from 0 to 40 and put values at the ends of bins into the right bin:

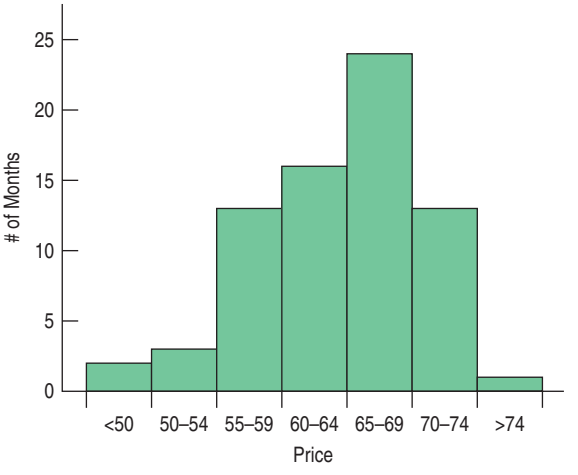
Downloads	Number OF Hours
0–5	2
5–10	2
10–15	3
15–20	5
20–25	6
25–30	3
30–35	2
35–40	1
Total	**24**

The histogram looks like this:



*Stem-and-Leaf Displays

Histograms provide an easy-to-understand summary of the distribution of a quantitative variable, but they don't show the data values themselves. **Stem-and-leaf displays** are like histograms, but they also give the individual values. They are easy to make by hand for data sets that aren't too large, so they're a great way to look at a small batch of values quickly.[1] Here's a stem-and-leaf display for the AIG stock data, alongside a histogram of the same data.



4|99
5|134
5|5667888999999
6|0000011122222344
6|555556666777788888889999
7|0001111222334
7|7

Figure 5.3 The AIG monthly average stock prices displayed both by a histogram (left) and stem-and-leaf display (right). Stem-and-leaf displays are typically made by hand, so we are most likely to use them for small data sets. For much larger data sets, we use a histogram.

- **How do stem-and-leaf displays work?** A stem-and-leaf display breaks each number into two parts: the stem shown to the left of the solid line and the leaf, to the right. For the AIG data, each price change, for example $67.02 is first truncated to two digits, $67. Then it is split into its two components: 6|7. The line 5|134 therefore, shows the values $51, $53, and $54 and corresponds to the histogram bin from $50 to $55. The stem-and-leaf in Figure 5.3 uses a bin width of 5. Another choice would be to increase the bin size and put all the prices from $50 to $60 on one line:

  5|1345667888999999

  That would decrease the number of bins to 4, but makes the bin from $60 to $70 too crowded:

  4|99
  5|1345667888999999
  6|000001112222344555556666777788888889999
  7|00011112223347

  Sometimes the stem-and-leaf display puts the higher numbers on top:

  7|7
  7|00001111222334
  6|555556666777788888889999
  6|0000011122222344
  5|5667888999999
  5|134
  4|99

  Either choice is possible, although putting the lower numbers on top makes the correspondence between histogram and stem-and-leaf easier to see.

[1]The authors like to make stem-and-leaf displays whenever data are presented (without a suitable display) at committee meetings or working groups. The insights from just that quick look at the distribution are often quite valuable.

In Chapter 4, you learned to check the Categorical Data Condition before making a pie chart or a bar chart. Now, by contrast, before making a stem-and-leaf display, or a histogram, you need to check the **Quantitative Data Condition**: The data are values of a quantitative variable whose units are known.

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

5.2 Shape

Once you've displayed the distribution in a histogram or stem-and-leaf display, what can you say about it? When you describe a distribution, you should pay attention to three things: its **shape**, its **center**, and its **spread**.

We describe the shape of a distribution in terms of its modes, its symmetry, and whether it has any gaps or outlying values.

Mode

Does the histogram have a single, central hump (or peak) or several, separated humps? These humps are called **modes**.[2] Formally, the mode is the single, most frequent value, but we rarely use the term that way. Sometimes we talk about the mode as being the value of the variable at the center of this hump. The AIG stock prices have a single mode around \$65 (Figure 5.1). We often use modes to describe the shape of the distribution. A distribution whose histogram has one main hump, such as the one for the AIG stock prices, is called **unimodal**; distributions whose histograms have two humps are **bimodal**, and those with three or more are called **multimodal**. For example, here's a bimodal distribution.

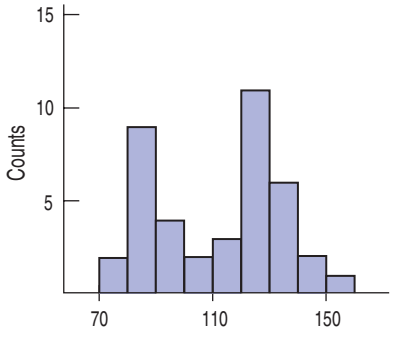


Figure 5.4 A bimodal distribution has two apparent modes.

A bimodal histogram is often an indication that there are two groups in the data. It's a good idea to investigate when you see bimodality.

A distribution whose histogram doesn't appear to have any mode and in which all the bars are approximately the same height is called **uniform**. (Chapter 9 gives a more formal definition.)

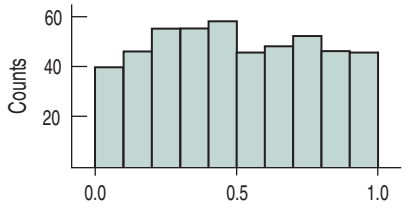


Figure 5.5 In a uniform distribution, bars are all about the same height. The histogram doesn't appear to have a mode.

[2]Technically, the mode is the value on the *x*-axis of the histogram below the highest peak, but informally we often refer to the peak or hump itself as a mode.

Where's the Mode?

The **mode** is typically defined as the single value that appears most often. That definition is fine for categorical variables because we need only to count the number of cases for each category. For quantitative variables, the meaning of *mode* is more ambiguous. For example, what's the mode of the AIG data? No two prices were exactly the same, but 7 months had prices between \$68 and \$69. Should that be the mode? Probably not—that seems a little arbitrary. For quantitative data, it makes more sense to use the word *mode* in the more general sense of "peak in a histogram," rather than as a single summary value.

Pie à la Mode

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

Symmetry

Could you fold the histogram along a vertical line through the middle and have the edges match pretty closely, as in Figure 5.6, or are more of the values on one side, as in the histograms in Figure 5.7? A distribution is **symmetric** if the halves on either side of the center look, at least approximately, like mirror images.

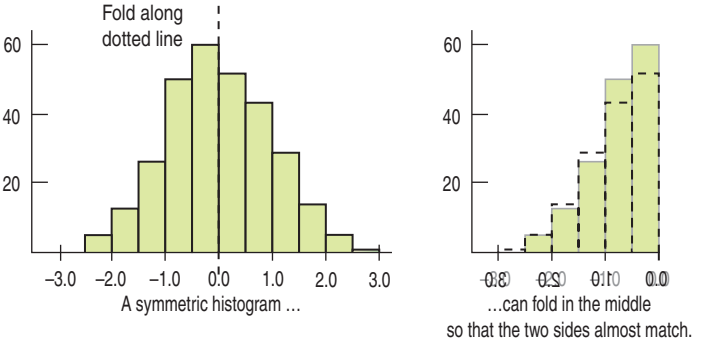


Figure 5.6 A symmetric histogram can fold in the middle so that the two sides almost match.

Amounts of things (dollars, employees, waiting times) can't be negative and have no natural upper limit. So, they often have distributions that are skewed to the right.

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

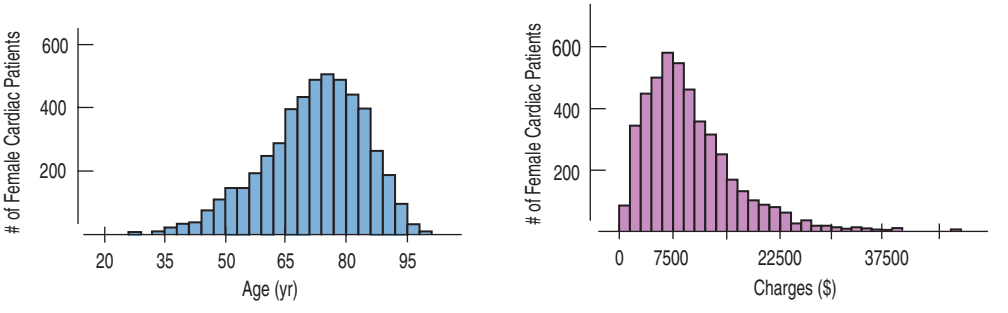


Figure 5.7 Two skewed histograms showing the age (left) and hospital charges (right) for all female heart attack patients in New York State in one year. The histogram of Age (in blue) is skewed to the left, while the histogram of Charges (in purple) is skewed to the right.

Outliers

Do any features appear to stick out? Often such features tell us something interesting or exciting about the data. You should always point out any stragglers or **outliers** that stand off away from the body of the distribution. For example, if you're studying the personal wealth of Americans and Bill Gates is in your sample, he would certainly be an outlier. Because his wealth would be so obviously atypical, you'd want to point it out as a special feature.

Outliers can affect almost every method we discuss in this book, so we'll always be on the lookout for them. An outlier can be the most informative part of your data, or it might just be an error. Either way, you shouldn't throw it away without comment. Treat it specially and discuss it when you report your conclusions about your data. (Or find the error and fix it if you can.) We'll soon learn a rule of thumb for how we can decide if and when a value might be considered to be an outlier and some advice for what to do when you encounter them.

- **Using Your Judgment.** How you characterize a distribution is often a judgment call. Does the gap you see in the histogram really reveal that you have two subgroups, or will it go away if you change the bin width slightly? Are those observations at the high end of the histogram truly unusual, or are they just the largest ones at the end of a long tail? These are matters of judgment on which different people can legitimately disagree. There's no automatic calculation or rule of thumb that can make the decision for you. Understanding your data and how they arose can help. What should guide your decisions is an honest desire to understand what is happening in the data.

  Looking at a histogram at several different bin widths can help you to see how persistent some of the features are. Some technologies offer ways to change the bin width interactively to get multiple views of the histogram. If the number of observations in each bin is small enough so that moving a couple of values to the next bin changes your assessment of how many modes there are, be careful. Be sure to think about the data, where they came from, and what kinds of questions you hope to answer from them.

For Example **Describing the shape of a distribution**

Question: Describe the shape of the distribution of downloads from the example on page 90.

Answer: It is symmetric and unimodal with no outliers.

5.3 Center

Look again at the AIG prices in Figure 5.1. If you had to pick one number to describe a *typical* price, what would you pick? When a histogram is unimodal and fairly symmetric, most people would point to the center of the distribution, where the histogram peaks. The typical price is around $65.00.

If we want to be more precise and *calculate* a number, we can *average* the data. In the AIG example, the average monthly prices is $64.48, about what we might expect from the histogram. You already know how to average values, but this is a good place to introduce notation that we'll use throughout the book. We'll call the generic variable y, and use the Greek capital letter sigma, Σ, to mean "sum" (sigma is "S" in Greek), and write[3]:

Notation Alert!

A bar over any symbol indicates the mean of that quantity.

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

According to this formula, we add up all the values of the variable, y, and divide that sum ($Total$, or Σy) by the number of data values, n. We call the resulting value the **mean** of y.[4]

[3] You may also see the variable called x and the equation written $\bar{x} = \frac{Total}{n} = \frac{\Sigma x}{n}$. We prefer to call a single variable y instead of x, because x will later be used to name a variable that predicts another (which we'll call y), but when you have only one variable either name is common. Most calculators call a single variable x.

[4] Once you've averaged the data, you might logically expect the result to be called the *average*. But average is used too colloquially as in the "average" home buyer, where we don't sum up anything. Even though average *is* sometimes used in the way we intend, as in the Dow Jones Industrial Average (which is actually a weighted average) or a batting average, we'll usually use the term *mean* throughout the book.

Although the mean is a natural summary for unimodal, symmetric distributions, it can be misleading for skewed data or for distributions with gaps or outliers. The histogram of AIG monthly prices in Figure 5.1 is unimodal, and nearly symmetric, with a slight left skew. A look at the total volume of stocks sold each month for the same 6 years tells a very different story. Figure 5.8 shows a unimodal but strongly right-skewed distribution with two gaps. The mean monthly volume was 170.1 million shares. Locate that value on the histogram. Does it seem a little high as a summary of a typical month's volume? In fact, more than two out of three months have volumes that are less than that value. It might be better to use the **median**—the value that splits the histogram into two equal *areas*. We find the median by counting in from the ends of the data until we reach the middle value. The median is commonly used for variables such as cost or income, which are likely to be skewed. That's because the median is *resistant* to unusual observations and to the shape of the distribution. For the AIG monthly trading volumes, the median is 135.9 million shares, which seems like a more appropriate summary.

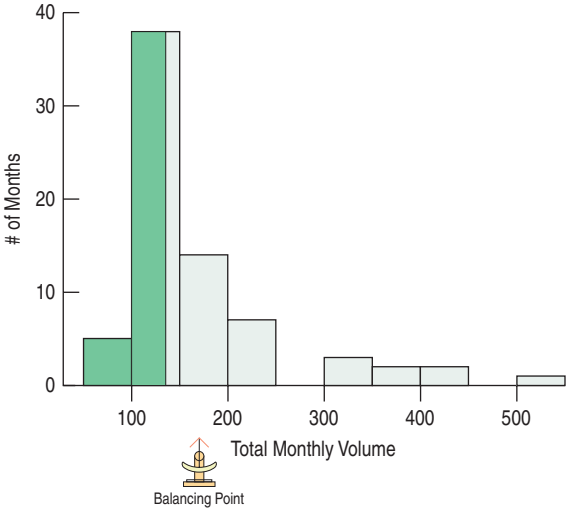


Figure 5.8 The median splits the area of the histogram in half at 135.9 million shares. Because the distribution is skewed to the right, the mean 170.1 million shares is *higher* than the median. The points at the right have pulled the mean toward them, away from the median.

Does it really make a difference whether we choose a mean or a median? The mean monthly price for the AIG stock is $64.48. Because the distribution of the prices is roughly symmetric, we'd expect the mean and median to be close. In fact, we compute the median to be $65.23. But for variables with skewed distributions, the story is quite different. For a right-skewed distribution like the monthly volumes in Figure 5.8, the mean is larger than the median: 170.1 compared to 135.9. The two give quite different summaries. The difference is due to the overall shape of the distributions.

By Hand

Finding the Median

Finding the median of a batch of n numbers is easy as long as you remember to order the values first. If n is odd, the median is the middle value.

Counting in from the ends, we find this value in the $\frac{n+1}{2}$ position.

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

Here are two examples:

Suppose the batch has the values 14.1, 3.2, 25.3, 2.8, −17.5, 13.9, and 45.8. First we order the values: −17.5, 2.8, 3.2, 13.9, 14.1, 25.3, and 45.8. There are 7 values, so the median is the $(7 + 1)/2 = 4$th value counting from the top or bottom: 13.9.

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

The mean is the point at which the histogram would balance. Just like a child who moves away from the center of a see-saw, a bar of the histogram far from the center has more leverage, pulling the mean in its direction. It's hard to argue that a summary that's been pulled aside by only a few outlying values or by a long tail is what we mean by the center of the distribution. That's why the median is usually a better choice for skewed data.

However, when the distribution is unimodal and symmetric, the mean offers better opportunities to calculate useful quantities and draw more interesting conclusions. It will be the summary value we work with much more throughout the rest of the book.

For Example **Finding the mean and median**

Question: From the data on page 90, what is a typical number of downloads per hour?

Answer: The mean number is 18.7 downloads per hour. The median is 19.5 downloads per hour. Because the distribution is unimodal and roughly symmetric, we shouldn't be surprised that the two are close. There are a few more hours (in the middle of the night) with small numbers of downloads that pull the mean lower than the median, but either one seems like a reasonable summary to report.

5.4 Spread of the Distribution

We know that the typical price of the AIG stock is around $65, but knowing the mean or median alone doesn't tell us about the entire distribution. A stock whose price doesn't move away from its center isn't very interesting.[5] The more the data vary, the less a measure of center can tell us. We need to know how spread out the data are as well.

One simple measure of spread is the **range**, defined as the difference between the extremes:

$$\text{Range} = max - min.$$

For the AIG price data, the range is $77.26 − $49.41 = $27.85. Notice that the range is *a single number* that describes the spread of the data, not an interval of values—as you might think from its use in common speech. If there are any unusual

[5] And not much of an investment, either.

observations in the data, the range is not resistant and will be influenced by them. Concentrating on the middle of the data avoids this problem.

The **quartiles** are the values that frame the middle 50% of the data. One quarter of the data lies below the lower quartile, Q1, and one quarter of the data lies above the upper quartile, Q3. The **interquartile range (IQR)** summarizes the spread by focusing on the middle half of the data. It's defined as the difference between the two quartiles:

$$\text{IQR} = \text{Q3} - \text{Q1}.$$

By Hand

Finding Quartiles

Quartiles are easy to find in theory, but more difficult in practice. The three quarties, Q1 (lower quartile), Q2 (the median) and Q3 (the upper quartile) split the sorted data values into quarters. So, for example, 25% of the data values will lie at or below Q1. The problem lies in the fact that unless your sample size divides nicely by 4, there isn't just one way to split the data into quarters. The statistical software package SAS offers at least five different ways to compute quartiles. The differences are usually small, but can be annoying. Here are two of the most common methods for finding quartiles by hand or with a calculator:

1. The Tukey Method

   Split the sorted data at the median. (If n is odd, include the median with each half). Then find the median of each of these halves—use these as the quartiles.

   Example: The data set {14.1, 3.2, 25.3, 2.8, −17.5, 13.9, 45.8}

   First we order the values: {−17.5, 2.8, 3.2, 13.9, 14.1, 25.3, 45.8}. We found the median to be 13.9, so form two data sets: {−17.5, 2.8, 3.2, 13.9} and { 13.9, 14.1, 25.3, 45.8}. The medians of these are $3.0 = (2.8 + 3.2)/2$ and $19.7 = (14.1 + 25.3)/2$. So we let $\text{Q1} = 3.0$ and $\text{Q3} = 19.7$.

2. The TI calculator method

   The same as the Tukey method, except we *don't* include the median with each half. So for {14.1, 3.2, 25.3, 2.8, −17.5, 13.9, and 45.8} we find the two data sets:

   {−17.5, 2.8, 3.2} and {14.1, 25.3, 45.8} by not including the median in either.

   Now the medians of these are $\text{Q1} = 2.8$ and $\text{Q3} = 25.3$.

   Notice the effect on the IQR. For Tukey:
   $\text{IQR} = \text{Q3} - \text{Q1} = 19.7 - 3.0 = 16.7$, but for TI,
   $\text{IQR} = 25.3 - 2.8 = 22.5$.

For both of these methods, notice that the quartiles are either data values, or the average of two adjacent values. In Excel, and other software, the quartiles are *interpolated*, so they may not be simple averages of two values. Be aware that there may be differences, but the idea is the same: the quartiles Q1, Q2, and Q3 split the data roughly into quarters.

For the AIG data, there are 36 values on either side of the median. After ordering the data, we average the 18th and 19th values to find $\text{Q1} = (59.96 + 60.26)/2 = \$60.11$. We average the 54th and 55th values to

Waiting in Line

Why do banks favor a single line that feeds several teller windows rather than separate lines for each teller? It does make the average waiting time slightly shorter, but that improvement is very small. The real difference people notice is that the time you can expect to wait is less variable when there is a single line, and people prefer consistency.

find Q3 = (68.99 + 69.02)/2 = \$69.01. So the IQR = Q3 − Q1 = \$69.01 − \$60.11 = \$8.90.

The IQR is usually a reasonable summary of spread, but because it uses only the two quartiles of the data, it ignores much of the information about how individual values vary.

A more powerful measure of spread—and the one we'll use most often—is the standard deviation, which, as we'll see, takes into account how far each value is from the mean. Like the mean, the standard deviation is appropriate only for symmetric data and can be influenced by outlying observations.

As the name implies, the standard deviation uses the *deviations* of each data value from the mean. If we tried to average these deviations, the positive and negative differences would cancel each other out, giving an average deviation of 0—not very useful. Instead, we square each deviation. The average[6] of the *squared* deviations is called the **variance** and is denoted by s^2:

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

The variance plays an important role in statistics, but as a measure of spread, it has a problem. Whatever the units of the original data, the variance is in *squared* units. We want measures of spread to have the same units as the data, so we usually take the square root of the variance. That gives the **standard deviation**.

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

For the AIG stock prices, s = \$6.12.

For Example — Describing the spread

Question: For the data on page 90, describe the spread of the number of downloads per hour.

Answer: The range of downloads is 36 − 2 = 34 downloads per hour.
The quartiles are 13 and 24.5, so the IQR is 24.5 − 13 = 11.5 downloads per hour. The standard deviation is 8.94 downloads per hour.

By Hand

Finding the Standard Deviation

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

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

Suppose the batch of values is 4, 3, 10, 12, 8, 9, and 3.

(continued)

[6]For technical reasons, we divide by $n - 1$ instead of n to take this average.

The mean is $\bar{y} = 7$. So find the deviations by subtracting 7 from each value:

Original Values	Deviations	Squared Deviations
4	$4 - 7 = -3$	$(-3)^2 = 9$
3	$3 - 7 = -4$	$(-4)^2 = 16$
10	$10 - 7 = 3$	9
12	$12 - 7 = 5$	25
8	$8 - 7 = 1$	1
9	$9 - 7 = 2$	4
3	$3 - 7 = -4$	16

Add up the squared deviations:
$9 + 16 + 9 + 25 + 1 + 4 + 16 = 80.$
Now, divide by $n - 1$: $80/6 = 13.33.$
Finally, take the square root: $s = \sqrt{13.33} = 3.65$

Just Checking

Thinking About Variation

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

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

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

5.5 Shape, Center, and Spread—A Summary

What should you report about a quantitative variable? Report the shape of its distribution, and include a center and a spread. But which measure of center and which measure of spread? The guidelines are pretty easy.

- If the shape is skewed, point that out and report the median and IQR. You may want to include the mean and standard deviation as well, explaining why the mean and median differ. The fact that the mean and median do not agree is a sign that the distribution may be skewed. A histogram will help you make the point.
- If the shape is unimodal and symmetric, report the mean and standard deviation and possibly the median and IQR as well. For unimodal symmetric data, the IQR is usually a bit larger than the standard deviation. If that's not true for your data set, look again to make sure the distribution isn't skewed or mutimodal and that there are no outliers.
- If there are multiple modes, try to understand why. If you can identify a reason for separate modes, it may be a good idea to split the data into separate groups.

- If there are any clearly unusual observations, point them out. If you are reporting the mean and standard deviation, report them computed with and without the unusual observations. The differences may be revealing.
- Always pair the median with the IQR and the mean with the standard deviation. It's not useful to report one without the other. Reporting a center without a spread can lead you to think you know more about the distribution than you do. Reporting only the spread omits important information.

For Example **Summarizing data**

Question: Report on the shape, center, and spread of the downloads data; see page 90.

Answer: The distribution of downloads per hour over the past 24 hours is unimodal and roughly symmetric. The mean number of downloads per hour is 18.7 and the standard deviation is 8.94. There are several hours in the middle of the night with very few downloads, but none seem to be so unusual as to be considered outliers.

5.6 Five-Number Summary and Boxplots

One good way to summarize a distribution with just a few values is with a five-number summary. The **five-number summary** of a distribution reports its median, quartiles, and extremes (maximum and minimum). For example, the five-number summary of the monthly trading volumes of AIG stock for the period 2002 to 2007 looks like this (in millions of shares).

Max	515.62
Q3	182.32
Median	135.87
Q1	121.04
Min	83.91

Table 5.2 The five-number summary of monthly trading volume of AIG shares (in millions of shares) for the period 2002 to 2007.

The five-number summary provides a good overall look at the distribution. For example, because the quartiles frame the middle half of the data, we can see that on half of the days the volume was between 121.04 and 182.32 million shares. We can also see the extremes of over 500 million shares on the high end and 83.91 million shares on the low end. Were those days extraordinary for some reason or just the busiest and quietest days? To answer that, we'll need to work with the summaries a bit more.

Once we have a five-number summary of a (quantitative) variable, we can display that information in a **boxplot** (see Figure 5.9).

A boxplot highlights several features of the distribution of a variable. The central box shows the middle half of the data, between the quartiles. Because the top of the box is at the third quartile (Q3) and the bottom is at Q1, the height of the box

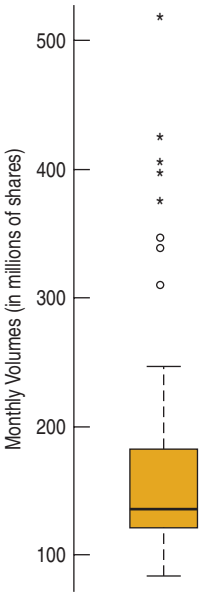


Figure 5.9 Boxplot of monthly volumes of AIG stock traded in the period 2002–2007 (in millions of shares).

is equal to Q3 − Q1 which is the IQR. (For the AIG data, it's 61.28.) The median is displayed as a horizontal line. If the median is roughly centered between the quartiles, then the middle half of the data is roughly symmetric. If it is not centered, the distribution is skewed. In extreme cases, the median can coincide with one of the quartiles.

The whiskers reach out from the box to the most extreme values that are not considered outliers. The boxplot nominates points as outliers if they fall farther than 1.5 IQRs beyond either quartile (for the AIG data, 1.5 IQR = 1.5 × 61.28 = 91.92). 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. This rule is not a definition of what makes a point an outlier. It just nominates cases for special attention. But it is not a substitute for careful analysis and thought about whether a value is special.

It's easy to make a boxplot. First locate the median and quartiles on an axis and draw three short lines. For the AIG data, those are at approximately 121 (Q1), 136 (median), and 182 (Q3). The axis is usually vertical (as in Figure 5.9), but it can be horizontal. Connect the quartile lines to make a box. Identify the "fences" at 1.5 IQR beyond each quartile. For the AIG data, that's 121.04 − 1.5 × 61.28 = 29.12 (lower fence) and 182.32 + 1.5 × 61.28 = 274.24 (upper fence). These fences are not drawn on the final boxplot. They are used to decide which points to display as outliers. Draw whiskers to the most extreme data value not outside the fences. In the AIG data, there are no values below the lower fence since the minimum 83.91 is greater than 29.12, but there are 7 points above the upper fence at 274.24. Finally, draw any outliers individually. Some boxplots use a special symbol for "far" outliers that lie more than 3 IQR's from the fences, as shown in Figure 5.9.

Some features of the distribution are lost in a boxplot, but as we'll soon see, they are especially useful when comparing several distributions side by side.

From the shape of the box in Figure 5.9, it looks like the central part of the distribution of volume is skewed to the right (upward here) and the dissimilar length of the two whiskers shows the outer parts of the distribution to be skewed as well. We also see several high volume and some extremely high volume days. Boxplots are particularly good at exhibiting outliers. These extreme days may deserve more attention. (When and why did they occur?)

The 1.5 IQR Rule for Nomination Outliers

Designate a point as an outlier if it lies farther than 1.5 IQRs from either the first (Q1) or third (Q3) quartile. Some boxplots also designate points as "far" outliers if they lie more than 3 IQRs from the quartiles. 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.

For Example — The boxplot rule for nominating outliers

Question: From the histogram on page 90, we saw that no points seemed to be so far from the center as to be considered outliers. Use the 1.5 IQR rule to see if it nominates any points as outliers.

Answer: The quartiles are 13 and 24.5 and the IQR is 11.5. 1.5 *IQR = 17.25. A point would have to be larger than 24.5 + 17.25 = 41.25 downloads/hr or smaller than 13 − 17.25 = −4.25. The largest value was 36 downloads/hr and all values must be nonnegative, so there are no points nominated as outliers.

Guided Example Credit Card Bank Customers

To focus on the needs of particular customers, companies often segment their customers into groups with similar needs or spending patterns. A major credit card bank wanted to see how much a particular group of cardholders charged per month on their cards in order to understand the potential growth in their card use. The data for each customer was the amount he or she spent using the card during a three-month period in 2008. Boxplots are especially useful for one variable when combined with a histogram and numerical summaries. Let's summarize the spending of this market segment.

PLAN

Setup Identify the *variable*, the time frame of the data, and the objective of the analysis.

We want to summarize the average monthly charges (in dollars) made by 500 cardholders from a market segment of interest during a three-month period in 2008. The data are quantitative, so we'll use histograms and boxplots, as well as numerical summaries.

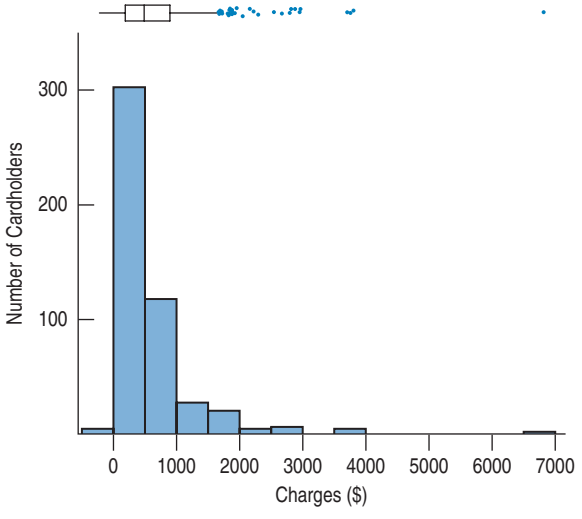
DO

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

REALITY CHECK

It is always a good idea to think about what you expected to see and to check whether the histogram is close to what you expected. Are the data about what you might expect for customers to charge on their cards in a month? A typical value is a few hundred dollars. That seems like the right ballpark.

Note that outliers are often easier to see with boxplots than with histograms, but the histogram provides more details about the shape of the distribution. This computer program "jitters" the outliers in the boxplot so they don't lie on top of each other, making them easier to see.



Both graphs show a distribution that is highly skewed to the right with several outliers and an extreme outlier near $7000.

Summary of Monthly Charges	
Count	500
Mean	544.749
Median	370.65
StdDev	661.244
IQR	624.125
Q1	114.54
Q3	738.665

The mean is much larger than the median. The data do not have a symmetric distribution.

(continued)

REPORT

Interpretation Describe the shape, center, and spread of the distribution. Be sure to report on the symmetry, number of modes, and any gaps or outliers.

Recommendation State a conclusion and any recommended actions or analysis.

MEMO

Re: Report on segment spending.

The distribution of charges for this segment during this time period is unimodal and skewed to the right. For that reason, we have summarized the data with the median and interquartile range (IQR).

The median amount charged was $370.65. Half of the cardholders charged between $114.54 and $738.67.

In addition, there are several high outliers, with one extreme value at $6745.

There are also a few negative values. We suspect that these are people who returned more than they charged in a month, but because the values might be data errors, we suggest that they be checked.

Future analyses should look at whether charges during these three months in 2008 were similar to charges in the rest of the year. We would also like to investigate if there is a seasonal pattern and, if so, whether it can be explained by our advertising campaigns or by other factors.

5.7 Comparing Groups

As we saw earlier, the volume of a stock can vary greatly from month to month or even day to day, but if we step back a bit, we may be able to find patterns that can help us understand, model, and predict it. We started the chapter by looking at monthly summaries of the price and volume of AIG stock. If, instead, we consider the individual daily values, we can group them into periods such as weeks, months, seasons, or years. The picture can change depending on what grouping we use. Comparing the distributions can reveal patterns, differences, and trends.

Let's start with the "big picture." Instead of taking monthly averages, let's look at the daily closing prices for the first two years of our data, 2002 and 2003:

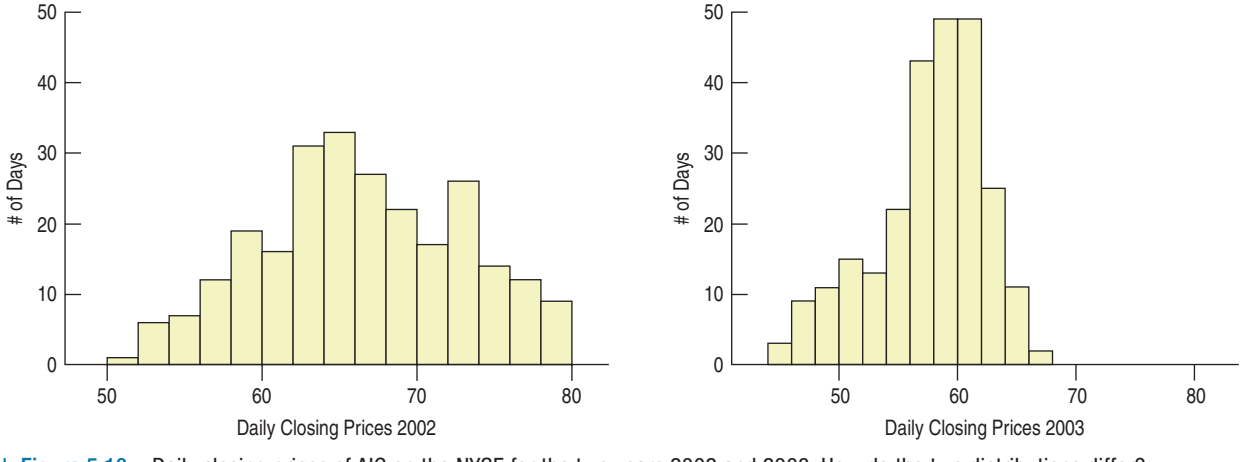


Figure 5.10 Daily closing prices of AIG on the NYSE for the two years 2002 and 2003. How do the two distributions differ?

It's not hard to see that prices were generally lower in 2003 than 2002. The price distribution for 2002 appears to be symmetric with a center in the high $60s while the 2003 distribution is left skewed with a center below $60. We were able to make the comparison easily because we displayed the two histograms on the same scale. Histograms with very different centers and spreads may appear similar unless you do that.

Histograms work well for comparing two groups, but what if we want to compare the prices across several years? Histograms are best at displaying one or two distributions. When we compare several groups, boxplots usually do a better job. Boxplots offer an ideal balance of information and simplicity, hiding the details while displaying the overall summary information. And we can plot them side by side, making it easy to compare multiple groups or categories.

When we place boxplots side by side, we can easily see which group has the higher median, which has the greater IQR, where the central 50% of the data is located, and which has the greater overall range. We can also get a general idea of symmetry from whether the medians are centered within their boxes and whether the whiskers extend roughly the same distance on either side of the boxes. Equally important, we can see past any outliers in making these comparisons because they've been displayed separately. We can also begin to look for trends in the medians and in the IQRs.

Guided Example AIG Stock Price and Volume

What really happened to the AIG stock price from the beginning of the period we've been studying through the financial crisis of 2008/2009? Boxplots of the number of shares traded by month are a good way to see such patterns. We're interested not only in the centers, but also in the spreads. Are volumes equally variable from year to year or are they more spread out in some years?

PLAN

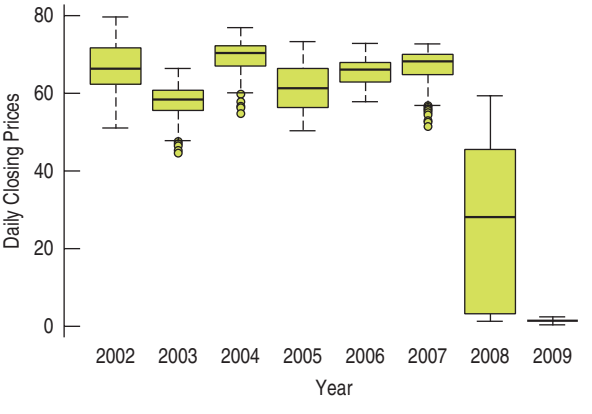
Setup Identify the variables, report the time frame of the data, and state the objective.

We want to compare the daily price of shares traded from year to year on the NYSE from 2002 through 2009.

The daily price is quantitative and measured in dollars. We can partition the values by year and use side-by-side boxplots to compare the daily prices across years.

DO

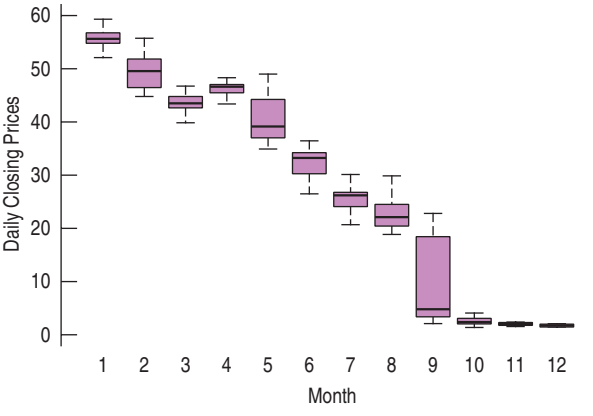
Mechanics Plot the side-by-side boxplots of the data.



(continued)

Display any other plots suggested by the previous.

What happened in 2008? We'd better look there with a finer partition. Here are boxplots by month for 2008.



REPORT

Conclusion Report what you've learned about the data and any recommended action or analysis.

MEMO

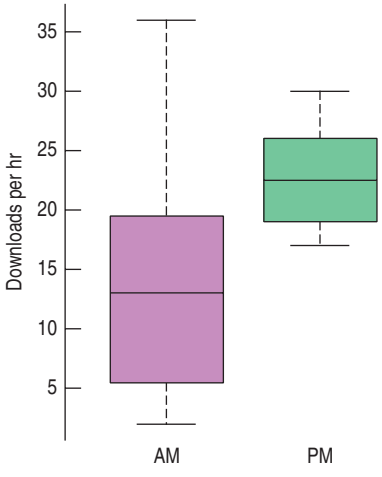
Re: Research on price of AIG stock

We have examined the daily closing prices of AIG stock on the NYSE for the period 2002 through 2009. As the displays show, prices were relatively stable for the period 2002 through 2007. Prices lowered in 2003 but recovered and stayed generally above $60 for 2004 through 2007. Then in 2008, prices dropped dramatically, and throughout 2009 AIG's stock price was a small fraction of what it had once been. A boxplot by month during 2008 shows that the decline in price was constant throughout the entire year but most noticeable in September 2008. Most analysts point to that month as the beginning of the financial meltdown, but clearly there were signs in the price of AIG that trouble had been brewing for much longer. By October and for the rest of the year, the price was very low with almost no variation.

For Example — Comparing boxplots

Question: For the data on page 90, compare the AM downloads to the PM downloads by displaying the two distributions side-by-side with boxplots.

Answer: There are generally more downloads in the afternoon than in the morning. The median number of afternoon downloads is around 22 as compared with 14 for the morning hours. The PM downloads are also much more consistent. The entire range of the PM hours, 15, is about the size of the IQR for AM hours. Both distributions appear to be fairly symmetric, although the AM hour distribution has some high points which seem to give some asymmetry.



5.8 Identifying Outliers

We've just seen that the price of AIG shares dropped precipitously during the year 2008. Let's look at a boxplot by month of the daily volumes to see if a similar pattern appears.

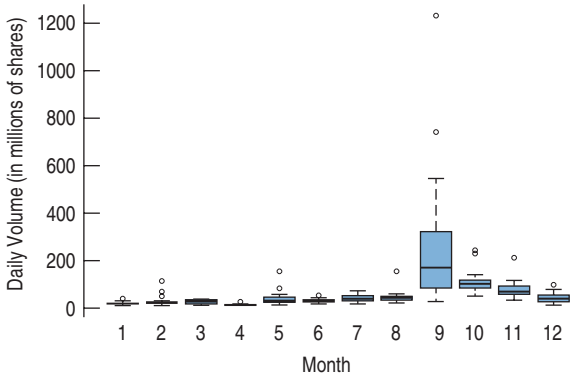


Figure 5.11 In January, there was a high volume day of 38 million shares that is nominated as an outlier for that month. In February there were three outliers with a maximum of over 100 million shares. In most months one or more high volume days are identified as outliers for their month. But none of these high volume days would have been considered unusual during September, when the median daily volume of AIG stock was 170 million shares. Days that may have seemed ordinary for September if placed in another month would have seemed extraordinary and *vice versa*. That high volume day in January certainly wouldn't stand out in September or even October or November, but for January it was remarkable.

Cases that stand out from the rest of the data deserve our attention. Boxplots have a rule for nominating extreme cases to display as outliers, but that's just a rule of thumb—not a definition. The rule doesn't tell you what to do with them. It's never a substitute for careful thinking about the data and their context.

So, what *should* we do with outliers? The first thing to do is to try to understand them in the context of the data. Once you've identified likely outliers, you should always investigate them. Some outliers are unbelievable and may simply be errors. A decimal point may have been misplaced, digits transposed, or digits repeated or omitted. Sometimes a number is transcribed incorrectly, perhaps copying an adjacent value on the original data sheet. Or, the units may be wrong. If you saw the number of AIG shares traded on the NYSE listed as 2 shares for a particular day, you'd know something was wrong. It could be that it was meant as 2 million shares, but you'd have to check to be sure. If you can identify the error, then you should certainly correct it.

Many outliers are not wrong; they're just different. These are the cases that often repay your efforts to understand them. You may learn more from the extraordinary cases than from summaries of the overall dataset.

What about those two days in September that stand out as extreme even during that volatile month? Those were September 15 and 16, 2008. On the 15th, 740 million shares of AIG stock were traded. That was followed by an incredible volume of over one billion shares of stock from a single company traded the following day. Here's how Barron's described the trading of September 16:

Record Volume for NYSE Stocks, Nasdaq Trades Surge Beats Its July Record

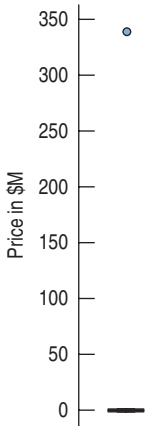
Yesterday's record-setting volume of 8.14 billion shares traded of all stocks listed on the New York Stock Exchange was pushed aside today by 9.31 billion shares in NYSE Composite volume. The biggest among those trades was the buying and selling of American International Group, with 1.11 billion shares traded as of 4 p.m. today. The AIG trades were 12% of all NYSE Composite volume.

For Example — Identifying outliers and summarizing data

Question: A real estate report lists the following prices for sales of single family homes in a small town in Virginia (rounded to the nearest thousand). Write a couple of sentences describing house prices in this town.

155,000	329,000	172,000	122,000	260,000
139,000	178,000	339,435,000	136,000	330,000
158,000	194,000	279,000	167,000	159,000
149,000	160,000	231,000	136,000	128,000

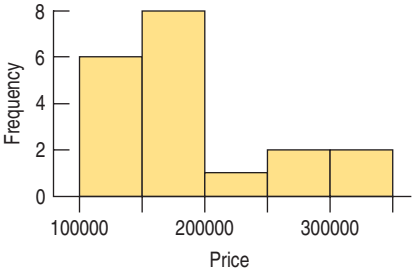
Answer: A box plot shows an extreme outlier:



That extreme point is a home whose sale price is listed at $339.4 M.

A check on the Internet shows that the most expensive homes ever sold are less than $200 M. This is clearly a mistake.

Setting aside this point, we find the following histogram and summary statistics:



The distribution of prices is strongly skewed to the right. The median price is $160,000. The minimum is $122,000 and the maximum (without the outlier) is $330,000. The middle 50% of house prices lie between $144,000 and $212,500 with an IQR of $68,500.

5.9 Standardizing

Forbes magazine lists the 258 largest privately held companies in the United States. What do we mean by large? *Forbes* provides two measures: *Revenue* (measured in $B) and number of *Employees*. *Forbes* uses only revenue in its rankings, but couldn't the size of the workforce be a way to measure size as well? How can we compare the two? How does having revenue of $20B compare to having a workforce of 50,000 employees? Which is "larger"? They don't have the same units, so we can't compare them directly. The trick is to standardize each variable first and then compare them. By doing this, we avoid comparing apples to oranges. Over and over during this course (and in many other courses you may take), questions such as "How does this value compare to a typical value?" or "How different are these two values?" will be answered by measuring the distance or difference in standard deviations from the mean.

Here are two companies listed by *Forbes*:

US Foodservice (A diversified food company, ranked #11) with $19.81B revenue and 26,000 employees

Toys "R" Us (the toy chain, ranked #21) with revenues of only $13.72B but 69,000 employees

It's easy to see which company earns more and which has more employees, but which company stands out more relative to others in the Forbes list?

How does standardizing work?

We first need to find the mean and standard deviation of each variable for all 258 companies in the *Forbes* list:

	Mean (all companies)	SD (all companies)
Revenue ($B)	6.23	10.56
Employees	19,629	32,055

Next we measure how *far* each of our values are by subtracting the mean and then dividing by the standard deviation:

$$z = (y - \bar{y})/s$$

We call the resulting value a **standardized value** and denote it with the letter z. Usually, we just call it a ***z*-score**. The z-score tells us how many standard deviations the value is from its mean.

Let's look at revenues first.

To compute the z-score for US Foodservice, take its value (19.81), subtract the mean (6.23) and divide by 10.56:

$$z = (19.81 - 6.23)/10.56 = 1.29$$

That means that US Foodservice's revenue is 1.29 standard deviations *above* the mean. How about employees?

$$z = (26{,}000 - 19{,}629)/32{,}055 = 0.20$$

So US Foodservice's workforce is not nearly as large (relative to the rest of the companies) as their revenue. The number of employees is only 0.20 standard deviations larger than the mean.

What about Toys "R" Us?

For revenue, $z = (13.72 - 6.23)/10.56 = 0.71$ and for employees, $z = (69{,}000 - 19{,}629)/32{,}055 = 1.54$

So who's bigger? If we use revenue, US Foodservice is the winner. If we use workforce, it's Toys "R" Us.

It's not clear which one we should use, but standardizing gives us a way to compare variables even when they're measured in different units. In this case, one could argue that Toys "R" Us is the bigger company. Its revenue z-score is 0.71 compared to US Foodservice's 1.29 but its employee size is 1.54 compared to 0.20 for US Foodservice.

It's not clear how to combine these two variables, although people do this sort of thing all the time. *Fortune* magazine with the help of the Great Places to Work Institute ranks the best companies to work for. In 2009 the software company SAS won. How did they get that honor? Overall, the analysts measured 50 different aspects of the companies. Was SAS better on all 50 variables? Certainly not, but it's almost certain that to combine the variables the analysts had to standardize the variables before combining them, no matter what their methodology.

Standardizing into *z*-Scores:

- Shifts the mean to 0.
- Changes the standard deviation to 1.
- Does not change the shape.
- Removes the units.

For Example — Comparing values by standardizing

Question: A real estate analyst finds more data from home sales as discussed in the example on page 106. Of 350 recent sales, the average price was \$175,000 with a standard deviation of \$55,000. The size of the houses (in square feet) averaged 2100 sq. ft. with a standard deviation of 650 sq. ft. Which is more unusual, a house in this town that costs \$340,000, or a 5000 sq. ft. house?

Answer: Compute the z-scores to compare. For the \$340,000 house:

$$z = \frac{y - \bar{y}}{s} = \frac{(340{,}000 - 175{,}000)}{55{,}000} = 3.0$$

The house price is 3 standard deviations above the mean.
For the 5000 sq. ft. house:

$$z = \frac{y - \bar{y}}{s} = \frac{(5{,}000 - 2{,}100)}{650} = 4.46$$

This house is 4.46 standard deviations above the mean in size. That's more unusual than the house that costs \$340,000.

5.10 Time Series Plots

The price and volume of stocks traded on the NYSE are reported daily. Earlier, we grouped the days into months and years, but we could simply look at the price day by day. A histogram can provide information about the distribution of a variable, but it can't show any pattern over time. Whenever we have time series data, it is a good idea to look for patterns by plotting the data in time order. Figure 5.12 shows the *daily prices* plotted over time for 2007.

A display of values against time is called a **time series plot**. This plot reflects the pattern that we were unable to see by displaying the entire year's prices in either a histogram or a boxplot. Now we can see that although the price rallied in the spring of 2007, after July there were already signs that the price might not stay above $60. By October, that pattern was clear.

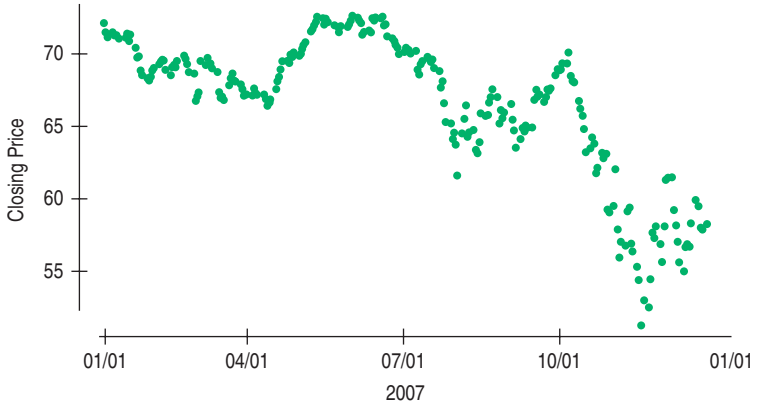


Figure 5.12 A time series plot of daily closing *prices* of AIG stock shows the overall pattern and changes in variation.

Time series plots often show a great deal of point-to-point variation, as Figure 5.12 does, and you'll often see time series plots drawn with all the points connected, especially in financial publications.

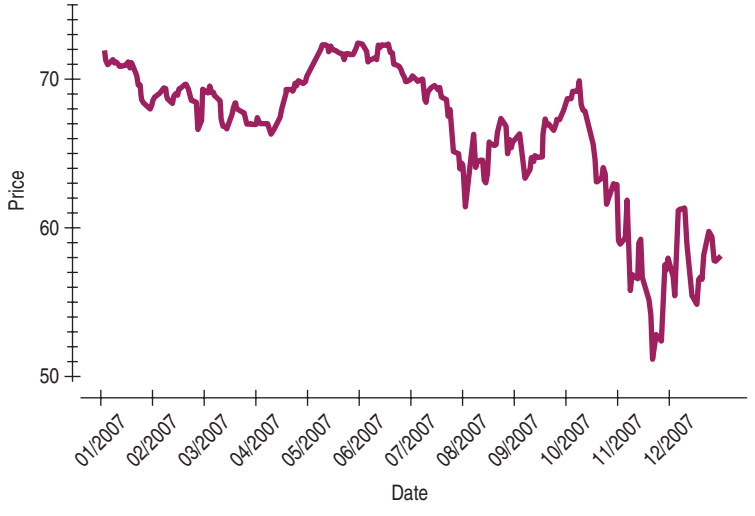


Figure 5.13 The *daily prices* of Figure 5.12, drawn by connecting all the points. Sometimes this can help us see the underlying pattern.

Often it is better to try to smooth out the local point-to-point variability. After all, we usually want to see past this variation to understand any underlying trend and think about how the values vary around that trend—the time series version of center and spread. There are many ways for computers to run a smooth trace through a time series plot. Some follow local bumps, others emphasize long-term trends. Some provide an equation that gives a typical value for any given time point, others just offer a smooth trace.

A smooth trace can highlight long-term patterns and help us see them through the more local variation. Figure 5.14 shows the daily prices of Figures 5.12 and 5.13 with a typical smoothing function, available in many statistics programs. With the smooth trace, it's a bit easier to see a pattern. The trace helps our eye follow the main trend and alerts us to points that don't fit the overall pattern.

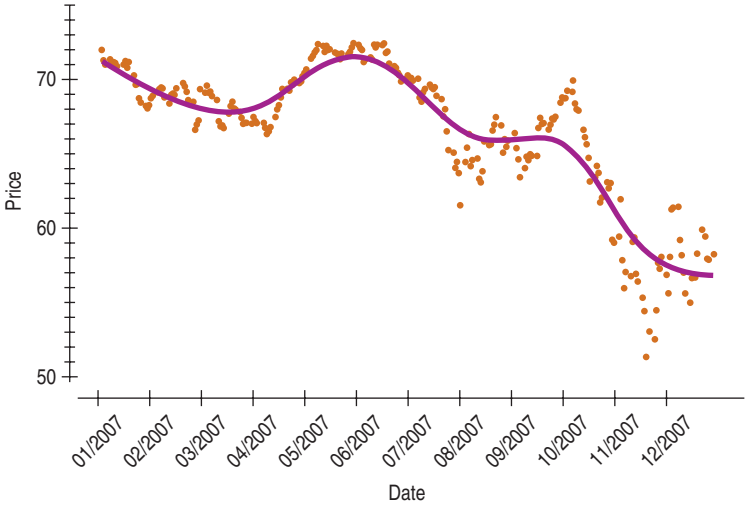


Figure 5.14 The daily volumes of Figure 5.12, with a smooth trace added to help your eye see the long-term pattern.

It is always tempting to try to extend what we see in a timeplot into the future. Sometimes that makes sense. Most likely, the NYSE volume follows some regular patterns throughout the year. It's probably safe to predict more volume on triple witching days (when contracts expire) and less activity in the week between Christmas and New Year's Day.

Other patterns are riskier to extend into the future. If a stock's price has been rising, how long 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 measures 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 indefinitely. 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.

Look at the prices in Figures 5.12 through 5.14 and try to guess what happened in the subsequent months. Was that drop from October to December a sign of trouble ahead, or was the increase in December back to around $60 where the stock had comfortably traded for several years a sign that stability had returned to AIG's

stock price? Perhaps those who picked up the stock for $51 in early November really got a bargain. Let's look ahead to 2008:

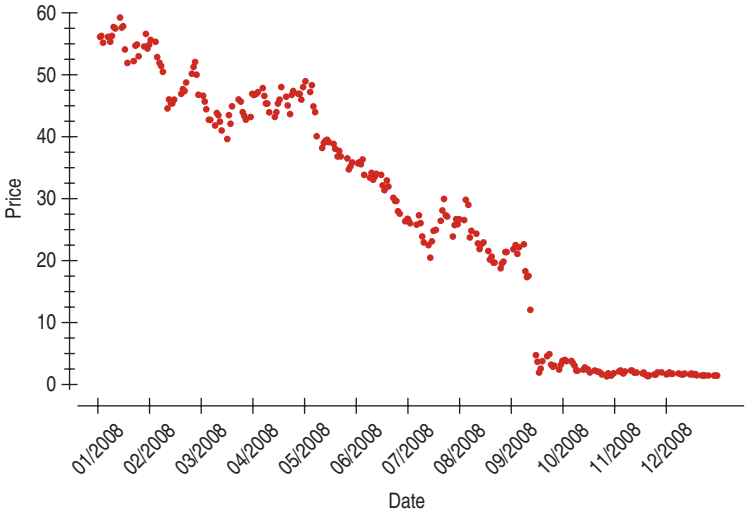


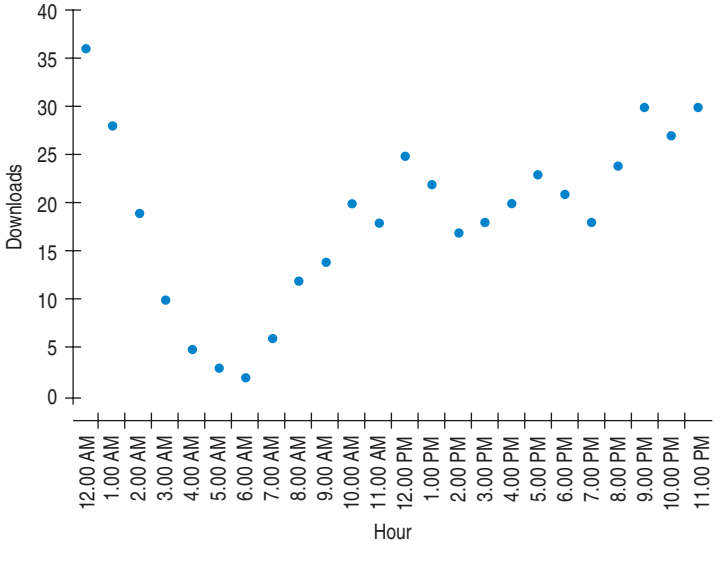
Figure 5.15 A time series plot of daily AIG *prices* shows what happened to the company in 2008.

Even through the spring of 2008, although the price was gently falling, nothing prepared traders following only the time series plot for what was to follow. In September the stock lost 99% of its value and as of 2010 was still trading below $2 an original share.

For Example — Plotting time series data

Question: The downloads from the example on page 90 are a time series. Plot the data by hour of the day and describe any patterns you see.

Answer: For this day, downloads were highest at midnight with about 36 downloads/hr, then dropped sharply until about 5–6 AM when they reached their minimum at 2–3 per hour. They gradually increased to about 20/hr by noon, and then stayed in the twenties until midnight, with a slight increase during the evening hours. When we split the data at midnight and noon, as we did earlier, we missed this pattern entirely.



The histogram we saw the beginning of the chapter (Figure 5.1) summarized the distribution of prices fairly well because during that period the prices were fairly stable. When a time series is **stationary**[7] (without a strong trend or change in variability), a histogram can provide a useful summary, especially in conjunction with a time series plot. However, when the time series is not stationary as was the case for AIG prices after 2007, a histogram is unlikely to capture much of interest. Then, a time series plot is the best graphical display to use in describing the behavior of the data.

5.11 Transforming Skewed Data

When a distribution is skewed, it can be hard to summarize the data 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 apply a simple function to each data value. One such function that can change the shape of a distribution is the logarithmic function. Let's examine an example in which a set of data is severely skewed.

In 1980, the average CEO made about 42 times the average worker's salary. In the two decades that followed, CEO compensation soared when compared with the average worker's pay; by 2000, that multiple had jumped to 525.[8] What does the distribution of the Fortune 500 companies' CEOs look like? Figure 5.16 shows a boxplot and a histogram of the 2005 compensation.

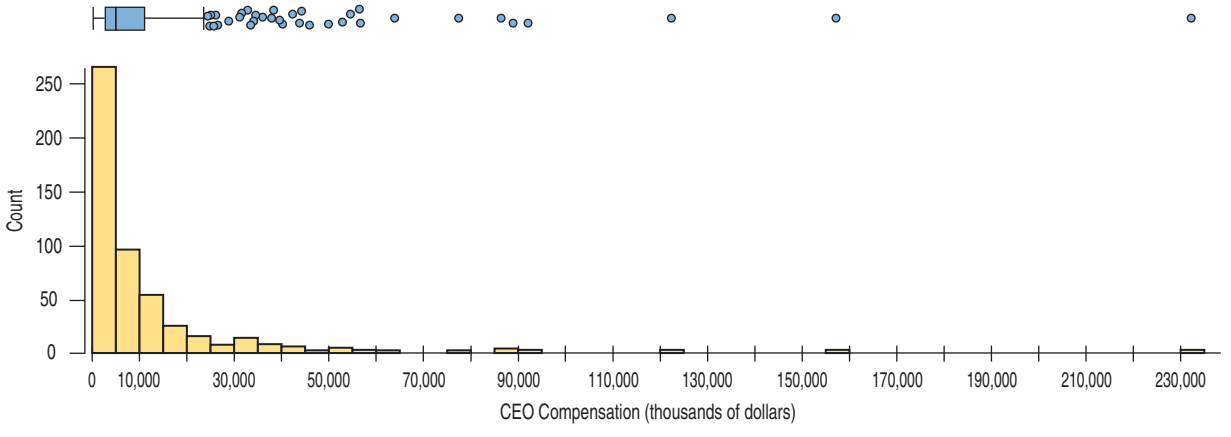


Figure 5.16 The total compensation for CEOs (in $000) of the 500 largest companies is skewed and includes some extraordinarily large values.

These values are reported in *thousands* of dollars. The boxplot indicates that some of the 500 CEOs received extraordinarily high compensation. The first bin of the histogram, containing about half the CEOs, covers the range $0 to $5,000,000. The reason that the histogram seems to leave so much of the area blank is that the largest observations are so far from the bulk of the data, as we can see from the boxplot. Both the histogram and boxplot make it clear that this distribution is very skewed to the right.

---

[7]Sometimes we separate out the properties and say the series is stationary with respect to the mean (if there is no trend) or stationary with respect to the variance (if the spread doesn't change), but unless otherwise noted, we'll assume that *all the statistical properties* of a stationary series are constant over time.

[8]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.

Total compensation for CEOs consists of their base salaries, bonuses, and extra compensation, usually in the form of stock or stock options. Data that add together several variables, such as the compensation data, can easily have skewed distributions. It's often a good idea to separate the component variables and examine them individually, but we don't have that information for the CEOs.

Skewed distributions are difficult to summarize. It's hard to know what we mean by the "center" of a skewed distribution, so it's not obvious what value to use 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 something different about how the data are distributed.

One way to make a skewed distribution more symmetric is to **re-express**, or **transform**, the data by applying a simple function to all the data values. Variables with a distribution that is skewed to the right often benefit from a re-expression by logarithms or square roots. Those skewed to the left may benefit from squaring the data values. It doesn't matter what base you use for a logarithm.

- **Dealing with logarithms** You probably don't encounter logarithms every day. In this book, we use them to make data behave better by making model assumptions more reasonable. 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. Fortunately, with technology, remaking a histogram or other display of the data is as easy as pushing a button.

The histogram of the logs of the total CEO compensations in Figure 5.17 is much more symmetric, so we can see that a typical *log compensation* is between 6.0 and 7.0, which means that it lies between $1 million and $10 million. To be more precise, the mean log_{10} value is 6.73, while the median is 6.67 (that's $5,370,317 and $4,677,351, respectively). Note that nearly all the values are between 6.0 and 8.0—in other words, between $1,000,000 and $100,000,000 per year. Logarithmic transformations are common, and because computers and calculators are available to do the calculating, you should consider transformation as a helpful tool whenever you have skewed data.

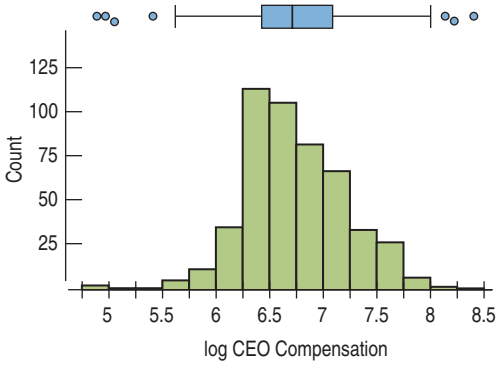
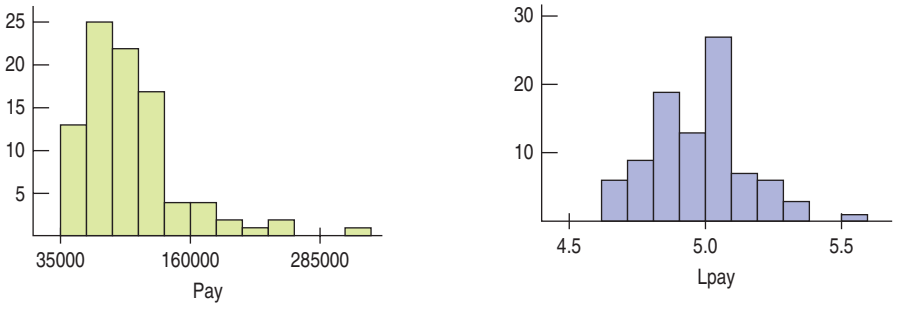


Figure 5.17 Taking logs makes the histogram of CEO total compensation nearly symmetric.

For Example — Transforming skewed data

Question: Every year *Fortune* magazine publishes a list of the 100 best companies to work for (http://money.cnn.com/magazines/fortune/bestcompanies/2010/). One statistic often looked at is the average annual pay for the most common job title at the company. Can we characterize those pay values? Here is a histogram of the average annual pay values and a histogram of the logarithm of the pay values. Which would provide the better basis for summarizing pay?



Answer: The pay values are skewed to the high end. The logarithm transformation makes the distribution more nearly symmetric. A symmetric distribution is more appropriate to summarize with a mean and standard deviation.

What Can Go Wrong?

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

The task of summarizing a quantitative variable requires that we follow a set of rules. We need to watch out for certain features of the data that make summarizing them with a number dangerous. Here's some advice:

- **Don't make a histogram of a categorical variable.** Just because the variable contains numbers doesn't mean it's quantitative. Here's a histogram of the insurance policy numbers of some workers. It's not very informative because the policy numbers are categorical. A histogram or stem-and-leaf display of a categorical variable makes no sense. A bar chart or pie chart may do better.

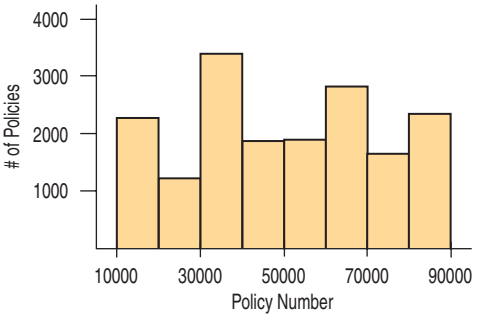


Figure 5.18 It's not appropriate to display categorical data like policy numbers with a histogram.

- **Choose a scale appropriate to the data.** Computer programs usually do a pretty good job of choosing histogram bin widths. Often, there's an easy way to adjust the width, sometimes interactively. Figure 15.19 shows the AIG price change histogram with two other choices for the bin size.
- **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 draw them on the same scale.
- **Label clearly.** Variables should be identified clearly and axes labeled so a reader knows what the plot displays.

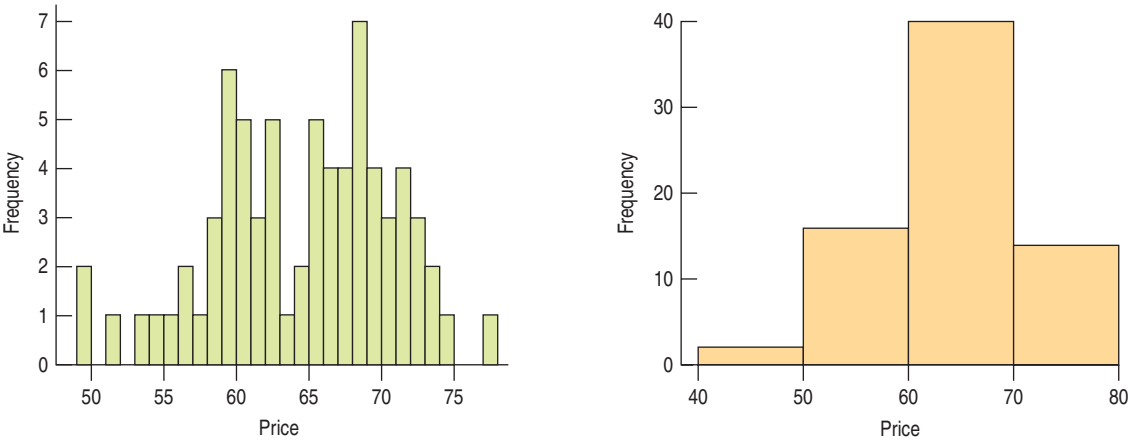


Figure 5.19 Changing the bin width changes how the histogram looks. The AIG stock prices look very different with these two choices.

Here's a remarkable example of a plot gone wrong. It illustrated a news story about rising college costs. It uses time series plots, 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.

(*continued*)

What's wrong? Just about everything.

- The horizontal scales are inconsistent. Both lines show trends over time, but 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 using two different scales. 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.

- **Do a reality check.** Don't let the computer (or calculator) do your thinking for you. Make sure the calculated summaries make sense. For example, does the mean look like it is in the center of the histogram? Think about the spread. An IQR of 50 mpg would clearly be wrong for a family car. And no measure of spread can be negative. The standard deviation can take the value 0, but only in the very unusual case that all the data values equal the same number. If you see the IQR or standard deviation equal to 0, it's probably a sign that something's wrong with the data.
- **Don't compute numerical summaries of a categorical variable.** The mean zip code or the standard deviation of Social Security numbers is not meaningful. If the variable is categorical, you should instead report summaries such as percentages. It is easy to make this mistake when you let technology do the summaries for you. After all, the computer doesn't care what the numbers mean.
- **Watch out for multiple modes.** If the distribution—as seen in a histogram, for example—has multiple modes, consider separating the data into groups. If you cannot separate the data in a meaningful way, you should not summarize the center and spread of the variable.
- **Beware of outliers.** If the data have outliers but are otherwise unimodal, consider holding the outliers out of the further calculations and reporting them individually. If you can find a simple reason for the outlier (for instance, a data transcription error), you should remove or correct it. If you cannot do either of these, then choose the median and IQR to summarize the center and spread.

Ethics in Action

Beth Tully owns Zenna's Café, an independent coffee shop located in a small midwestern city. Since opening Zenna's in 2002, she has been steadily growing her business and now distributes her custom coffee blends to a number of regional restaurants and markets. She operates a microroaster that offers specialty grade Arabica coffees recognized as some as the best in the area. In addition to providing the highest quality coffees, Beth also wants her business to be socially responsible. Toward that end, she pays fair prices to coffee farmers and donates funds to help charitable causes in Panama, Costa Rica, and Guatemala. In addition, she encourages her employees to get involved in the local community. Recently, one of the well-known multinational coffeehouse chains announced plans to locate shops in her area. This chain is

one of the few to offer Certified Free Trade coffee products and work toward social justice in the global community. Consequently, Beth thought it might be a good idea for her to begin communicating Zenna's socially responsible efforts to the public, but with an emphasis on their commitment to the local community. Three months ago she began collecting data on the number of volunteer hours donated by her employees per week. She has a total of 12 employees, of whom 10 are full time. Most employees volunteered less than 2 hours per week, but Beth noticed that one part-time employee volunteered more than 20 hours per week. She discovered that her employees collectively volunteered an average of 15 hours per month (with a median of 8 hours). She planned to report the average number and believed most people would be impressed with Zenna's level of commitment to the local community.

ETHICAL ISSUE *The outlier in the data affects the average in a direction that benefits Beth Tully and Zenna's Café (related to Item C, ASA Ethical Guidelines).*

ETHICAL SOLUTION *Beth's data are highly skewed. There is an outlier value (for a part-time employee) that pulls the average number of volunteer hours up. Reporting the average is misleading. In addition, there may be justification to eliminate the value since it belongs to a part-time employee (10 of the 12 employees are full time). It would be more ethical for Beth to: (1) report the average but discuss the outlier value, (2) report the average for only full-time employees, or (3) report the median instead of the average.*

What Have We Learned?

Learning Objectives

- Make and interpret histograms to display the distribution of a variable.
  - We understand distributions in terms of their shape, center, and spread.
- Describe the shape of a distribution.
  - A **symmetric** distribution has roughly the same shape reflected around the center
  - A **skewed** distribution extends farther on one side than on the other.
  - A **unimodal** distribution has a single major hump or mode; a bimodal distribution has two; multimodal distributions have more.
  - **Outliers** are values that lie far from the rest of the data.
- Compute the mean and median of a distribution, and know when it is best to use each to summarize the center.
  - The **mean** is the sum of the values divided by the count. It is a suitable summary for unimodal, symmetric distributions.
  - The **median** is the middle value; half the values are above and half are below the median. It is a better summary when the distribution is skewed or has outliers.
- Compute the standard deviation and interquartile range (IQR), and know when it is best to use each to summarize the spread.
  - The **standard deviation** is roughly the square root of the average squared difference between each data value and the mean. It is the summary of choice for the spread of unimodal, symmetric variables.
  - The **IQR** is the difference between the quartiles. It is often a better summary of spread for skewed distributions or data with outliers.
- Find a five-number summary and, using it, make a boxplot. Use the boxplot's outlier nomination rule to identify cases that may deserve special attention.
  - A **five-number summary** consists of the median, the quartiles, and the extremes of the data.
  - A **boxplot** shows the quartiles as the upper and lower ends of a central box, the median as a line across the box, and "whiskers" that extend to the most extreme values that are not nominated as outliers.

(continued)

- Boxplots display separately any case that is more than 1.5 IQRs beyond each quartile. These cases should be considered as possible outliers.

■ Use boxplots to compare distributions.

- Boxplots facilitate comparisons of several groups. It is easy to compare centers (medians) and spreads (IQRs).
- Because boxplots show possible outliers separately, any outliers don't affect comparisons.

■ Standardize values and use them for comparisons of otherwise disparate variables.

- We standardize by finding ***z*-scores**. To convert a data value to its z-score, subtract the mean and divide by the standard deviation.
- z-scores have no units, so they can be compared to z-scores of other variables.
- The idea of measuring the distance of a value from the mean in terms of standard deviations is a basic concept in Statistics and will return many times later in the course.

■ Make and interpret time plots for time series data.

- Look for the trend and any changes in the spread of the data over time.

Terms

Bimodal Distributions with two modes.

Boxplot A boxplot displays the 5-number summary as a central box with whiskers that extend to the nonoutlying values. Boxplots are particularly effective for comparing groups.

Center The middle of the distribution, usually summarized numerically by the mean or the median.

Distribution The distribution of a variable gives:

- possible values of the variable
- frequency or relative frequency of each value

Five-number summary A five-number summary for a variable consists of:

- The minimum and maximum
- The quartiles Q1 and Q3
- The median

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

Interquartile range (IQR) The difference between the first and third quartiles. IQR = Q3 − Q1.

Mean A measure of center found as $\bar{y} = \Sigma y/n$.

Median The middle value with half of the data above it and half below it.

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

Multimodal Distributions with more than two modes.

Outliers Extreme values that don't appear to belong with the rest of the data. They may be unusual values that deserve further investigation or just mistakes; there's no obvious way to tell.

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

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

Re-express or transform To re-express or transform data, take the logarithm, square root, reciprocal, or some other mathematical operation on all values of the data set. Re-expression can make the distribution of a variable more nearly symmetric and the spread of groups more nearly alike.

Shape	The visual appearance of the distribution. To describe the shape, look for: • single vs. multiple modes • symmetry vs. skewness
Skewed	A distribution is skewed if one tail stretches out farther than the other.
Spread	The description of how tightly clustered the distribution is around its center. Measures of spread include the IQR and the standard deviation.
Standard deviation	A measure of spread found as $s = \sqrt{\frac{\Sigma(y - \bar{y})^2}{n - 1}}$.
Standardized value	We standardize a value by subtracting the mean and dividing by the standard deviation for the variable. These values, called z-scores, have no units.
Stationary	A time series is said to be stationary if its statistical properties don't change over time.
Stem-and-leaf display	A stem-and-leaf display shows quantitative data values in a way that sketches the distribution of the data. It's best described in detail by example.
Symmetric	A distribution is symmetric if the two halves on either side of the center look approximately like mirror images of each other.
Tail	The tails of a distribution are the parts that typically trail off on either side.
Time series plot	Displays data that change over time. Often, successive values are connected with lines to show trends more clearly.
Uniform	A distribution that's roughly flat is said to be uniform.
Unimodal	Having one mode. This is a useful term for describing the shape of a histogram when it's generally mound-shaped.
Variance	The standard deviation squared.
z-score	A standardized value that tells how many standard deviations a value is from the mean; z-scores have a mean of 0 and a standard deviation of 1.

Technology Help: Displaying and Summarizing Quantitative Variables

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

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

```
Variable: Weight
N = 234
Mean =143.3       Median = 139
St. Dev = 11.1    IQR = 14
```

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

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

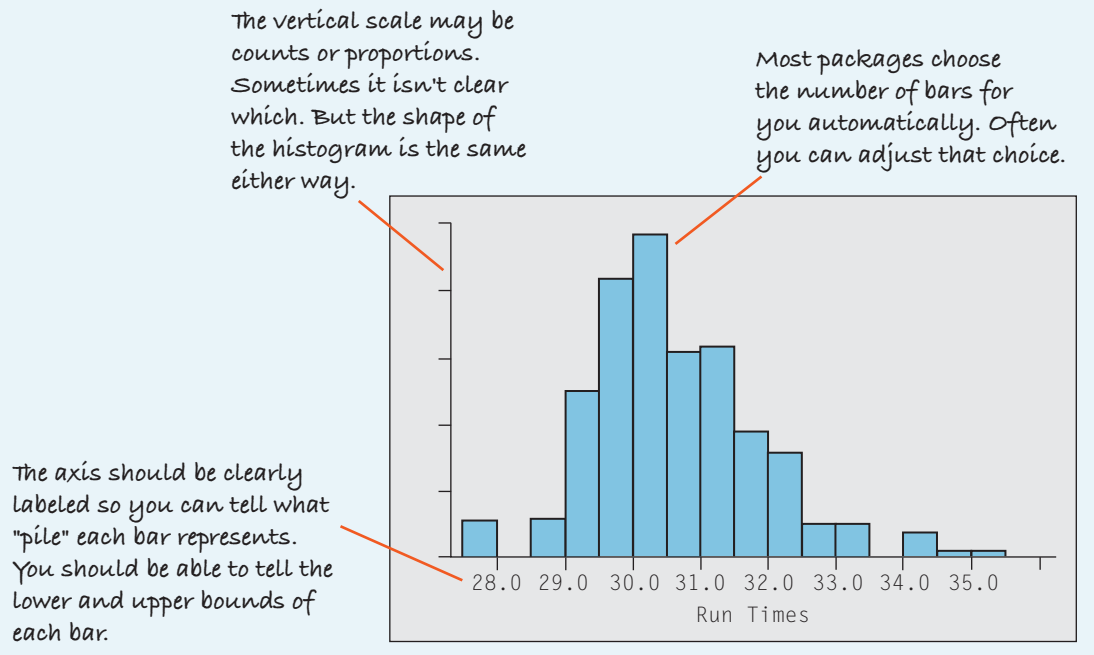
It is usually easy to read the results and identify each computed summary. You should be able to read the summary statistics produced by any computer package.

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

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

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

(continued)

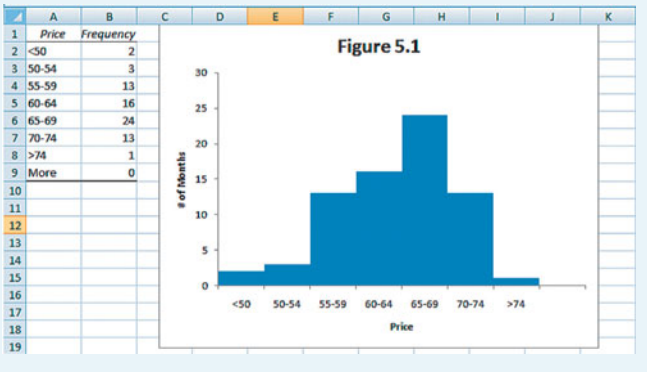


EXCEL

XLSTAT

To make a histogram in Excel 2007 or 2010, use the Data Analysis add-in. If you have not installed that, you must do that first.

- From the Data ribbon, select the Data Analysis add-in.
- From its menu, select Histograms.
- Indicate the range of the data whose histogram you wish to draw.
- Indicate the bin ranges that are up to and including the right end points of each bin.
- Check **Labels** if your columns have names in the first cell.
- Check **Chart output** and click **OK**.
- Right-click on any bar of the resulting graph and, from the menu that drops down, select **Format Data Series . . .**
- In the dialog box that opens, select **Series Options** from the sidebar.
- Slide the Gap Width slider to **No Gap**, and click **Close**.
- In the pivot table on the left, use your pointing tool to slide the bottom of the table up to get rid of the "more" bin.
- Edit the bin names in Column A to properly identify the contents of each bin.



- You can right click on the legend or axis names to edit or remove them.
- Following these instructions, you can reproduce Figure 5.1 using the data set AIG.

Alternatively, you can set up your own bin boundaries and count the observations falling within each bin using an Excel function such as FREQUENCY (Data array, Bins array). Consult your Excel manual or help files for details of how to do this.

JMP

To make a histogram and find summary statistics:

- Choose **Distribution** from the **Analyze** menu.
- In the **Distribution** dialog, drag the name of the variable that you wish to analyze into the empty window beside the label "**Y, Columns**."
- Click **OK**. JMP computes standard summary statistics along with displays of the variables.

To make boxplots:

- Choose **Fit y By x**. Assign a continuous response variable to **Y, Response** and a nominal group variable holding the group names to **X, Factor,** and click **OK**. JMP will offer (among other things) dotplots of the data. click the red triangle and, under **Display Options**, select Boxplots. Note: If the variables are of the wrong type, the display options might not offer boxplots.

MINITAB

To make a histogram:

- Choose **Histogram** from the **Graph** menu.
- Select "Simple" for the type of graph and click **OK**.
- Enter the name of the quantitative variable you wish to display in the box labeled "Graph variables." Click **OK**.

To make a boxplot:

- Choose **Boxplot** from the **Graph** menu and specify your data format.

To calculate summary statistics:

- Choose **Basic Statistics** from the **Stat** menu. From the **Basic Statistics** submenu, choose **Display Descriptive Statistics**.
- Assign variables from the variable list box to the Variables box. MINITAB makes a Descriptive Statistics table.

SPSS

To make a histogram or boxplot in SPSS open the Chart Builder from the Graphs menu.

- Click the **Gallery** tab.
- Choose **Histogram** or **Boxplot** from the list of chart types.
- Drag the icon of the plot you want onto the canvas.
- Drag a scale variable to the y-axis drop zone.
- Click **OK**.

To make side-by-side boxplots, drag a categorical variable to the x-axis drop zone and click **OK**.

To calculate summary statistics:

- Choose **Explore** from the **Descriptive Statistics** submenu of the **Analyze** menu. In the Explore dialog, assign one or more variables from the source list to the Dependent List and click the **OK** button.

Hotel Occupancy Rates

Many properties in the hospitality industry experience strong seasonal fluctuations in demand. To be successful in this industry it is important to anticipate such fluctuations and to understand demand patterns. The file **Occupancy_Rates** contains data on monthly *Hotel Occupancy Rates* (in % capacity) for Honolulu, Hawaii, from January 2000 to December 2007.

Examine the data and prepare a report for the manager of a hotel chain in Honolulu on patterns in *Hotel Occupancy* during this period. Include both numerical summaries and graphical displays and summarize the patterns that you see. Discuss any unusual features of the data and explain them if you can, including a discussion of whether the manager should take these features into account for future planning.

Value and Growth Stock Returns

Investors in the stock market have choices of how aggressive they would like to be with their investments. To help investors, stocks are classified as "growth" or "value" stocks. Growth stocks are generally shares in high quality companies that have demonstrated consistent performance and are expected to continue to do well. Value stocks on the other hand are stocks whose prices seem low compared to their inherent worth (as measured by the book to price ratio). Managers invest in these hoping that their low price is simply an overreaction to recent negative events.

In the data set **Returns**[9] are the monthly returns of 2500 stocks classified as Growth and Value for the time period January 1975 to June 1997. Examine the distributions of the two types of stocks and discuss the advantages and disadvantages of each. Is it clear which type of stock offers the best investment? Discuss briefly.

[9]Source: Independence International Associates, Inc. maintains a family of international style indexes covering 22 equity markets. The highest book-to-price stocks are selected one by one from the top of the list. The top half of these stocks become the constituents of the "value index," and the remaining stocks become the "growth index."

Exercises

SECTION 5.1

1. As part of the marketing team at an Internet music site, you want to understand who your customers are. You send out a survey to 25 customers (you use an incentive of $50 worth of downloads to guarantee a high response rate) asking for demographic information. One of the variables is the customer's age. For the 25 customers the ages are:

20	32	34	29	30
30	30	14	29	11
38	22	44	48	26
25	22	32	35	32
35	42	44	44	48

a) Make a histogram of the data using a bar width of 10 years.
b) Make a histogram of the data using a bar width of 5 years.
c) Make a relative frequency histogram of the data using a bar width of 5 years.
d) *Make a stem-and-leaf plot of the data using 10s as the stems and putting the youngest customers on the top of the plot.

2. As the new manager of a small convenience store, you want to understand the shopping patterns of your customers. You randomly sample 20 purchases from yesterday's records (all purchases in U.S. dollars):

39.05	2.73	32.92	47.51
37.91	34.35	64.48	51.96
56.95	81.58	47.80	11.72
21.57	40.83	38.24	32.98
75.16	74.30	47.54	65.62

a) Make a histogram of the data using a bar width of $20.
b) Make a histogram of the data using a bar width of $10.
c) Make a relative frequency histogram of the data using a bar width of $10.
d) *Make a stem-and-leaf plot of the data using $10 as the stems and putting the smallest amounts on top.

SECTION 5.2

3. For the histogram you made in Exercise 1a,

a) Is the distribution unimodal or multimodal?
b) Where is (are) the mode(s)?
c) Is the distribution symmetric?
d) Are there any outliers?

4. For the histogram you made in Exercise 2a:

a) Is the distribution unimodal or multimodal?
b) Where is (are) the mode(s)?
c) Is the distribution symmetric?
d) Are there any outliers?

SECTION 5.3

5. For the data in Exercise 1:

a) Would you expect the mean age to be smaller than, bigger than, or about the same size as the median? Explain.
b) Find the mean age.
c) Find the median age.

6. For the data in Exercise 2:

a) Would you expect the mean purchase to be smaller than, bigger than, or about the same size as the median? Explain.
b) Find the mean purchase.
c) Find the median purchase.

SECTION 5.4

7. For the data in Exercise 1:

a) Find the quartiles using your calculator.
b) Find the quartiles using the method on page 96.
c) Find the IQR using the quartiles from part b.
d) Find the standard deviation.

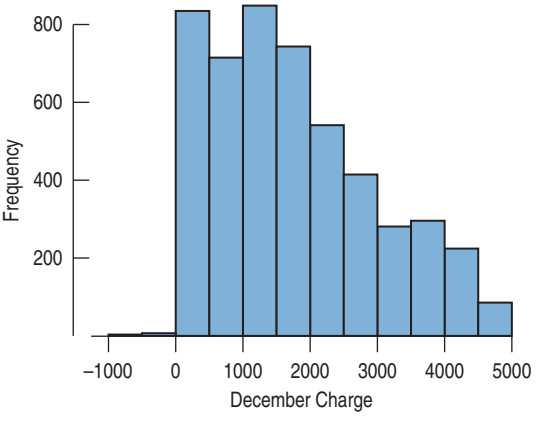
8. For the data in Exercise 2:

a) Find the quartiles using your calculator.
b) Find the quartiles using the method on page 96.
c) Find the IQR using the quartiles from part b.
d) Find the standard deviation.

SECTION 5.5

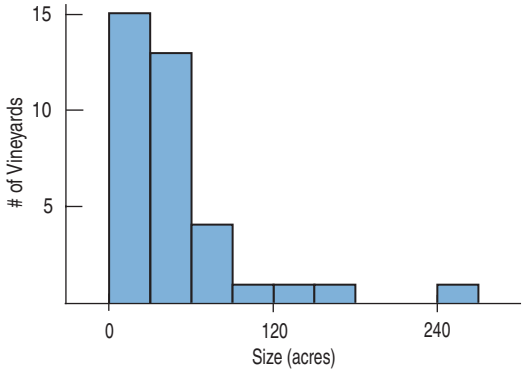
9. The histogram shows the December charges (in $) for 5000 customers from one marketing segment from a credit card company. (Negative values indicate customers who received more credits than charges during the month.)

a) Write a short description of this distribution (shape, center, spread, unusual features).
b) Would you expect the mean or the median to be larger? Explain.
c) Which would be a more appropriate summary of the center, the mean or the median? Explain.



10. Adair Vineyard is a 10-acre vineyard in New Paltz, New York. The winery itself is housed in a 200-year-old historic Dutch barn, with the wine cellar on the first floor and the tasting room and gift shop on the second. Since they are relatively small and considering an expansion, they are curious about how their size compares to that of other vineyards. The histogram shows the sizes (in acres) of 36 wineries in upstate New York.

a) Write a short description of this distribution (shape, center, spread, unusual features).
b) Would you expect the mean or the median to be larger? Explain.
c) Which would be a more appropriate summary of the center, the mean or the median? Explain.



SECTION 5.6

11. For the data in Exercise 1:

a) Draw a boxplot using the quartiles from Exercise 7b.
b) Does the boxplot nominate any outliers?
c) What age would be considered a high outlier?

12. For the data in Exercise 2:

a) Draw a boxplot using the quartiles from Exercise 8b.
b) Does the boxplot nominate any outliers?
c) What purchase amount would be considered a high outlier?

13. Here are summary statistics for the sizes (in acres) of upstate New York vineyards from Exercise 10.

Variable	N	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Acres	36	46.50	47.76	6	18.50	33.50	55	250

a) From the summary statistics, would you describe this distribution as symmetric or skewed? Explain.
b) From the summary statistics, are there any outliers? Explain.
c) Using these summary statistics, sketch a boxplot. What additional information would you need to complete the boxplot?

14. 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 these questions.

	% 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?
b) Are there any outliers? Explain.
c) Create a boxplot of these data.

SECTION 5.7

15. The survey from Exercise 1 had also asked the customers to say whether they were male or female. Here are the data:

Age	Sex	Age	Sex	Age	Sex	Age	Sex	Age	Sex
20	M	32	F	34	F	29	M	30	M
30	F	30	M	14	M	29	M	11	M
38	F	22	M	44	F	48	F	26	F
25	M	22	M	32	F	35	F	32	F
35	F	42	F	44	F	44	F	48	F

Construct boxplots to compare the ages of men and women and write a sentence summarizing what you find.

16. The store manager from Exercise 2 has collected data on purchases from weekdays and weekends. Here are some summary statistics (rounded to the nearest dollar):

Weekdays n = 230

Min = 4, Q1 = 28, Median = 40, Q3 = 68, Max = 95

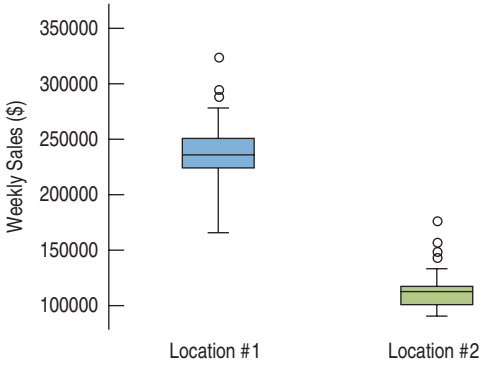
Weekend n = 150

Min = 10, Q1 = 35, Median = 55, Q3 = 70, Max = 100

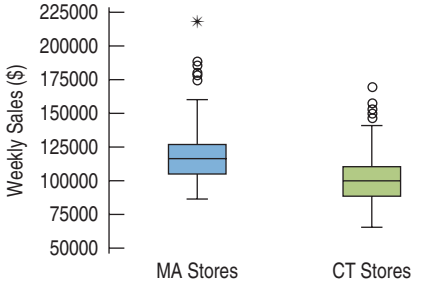
From these statistics, construct side-by-side boxplots and write a sentence comparing the two distributions.

17. Here are boxplots of the weekly sales (in $ U.S.) over a two-year period for a regional food store for two locations. Location #1 is a metropolitan area that is known to be residential where shoppers walk to the store. Location #2 is a suburban area where shoppers drive to the store. Assume that the two towns have similar populations and

that the two stores are similar in square footage. Write a brief report discussing what these data show.



18. Recall the distributions of the weekly sales for the regional stores in Exercise 17. Following are boxplots of weekly sales for this same food store chain for three stores of similar size and location for two different states: Massachusetts (MA) and Connecticut (CT). Compare the distribution of sales for the two states and describe in a report.



SECTION 5.9

19. Using the ages from Exercise 1:

a) Standardize the minimum and maximum ages using the mean from Exercise 5b and the standard deviation from Exercise 7d.
b) Which has the more extreme z-score, the min or the max?
c) How old would someone with a z-score of 3 be?

20. Using the purchases from Exercise 2:

a) Standardize the minimum and maximum purchase using the mean from Exercise 6b and the standard deviation from Exercise 8d.
b) Which has the more extreme z-score, the min or the max?
c) How large a purchase would a purchase with a z-score of 3.5 be?

SECTION 5.11

21. When analyzing data on the number of employees in small companies in one town, a researcher took square roots of the counts. Some of the resulting values, which are reasonably symmetric, were:

4, 4, 6, 7, 7, 8, 10

What were the original values, and how are they distributed?

22. You wish to explain to your boss what effect taking the base-10 logarithm of the salary values in the company's database will have on the data. As simple, example values, you compare a salary of $10,000 earned by a part-time shipping clerk, a salary of $100,000 earned by a manager, and the CEO's $1,000,000 compensation package. Why might the average of these values be a misleading summary? What would the logarithms of these three values be?

CHAPTER EXERCISES

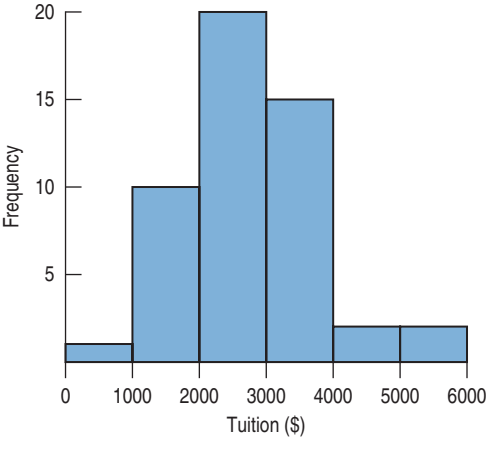
23. **Statistics in business.** Find a histogram that shows the distribution of a variable in a business publication (e.g., *The Wall Street Journal*, *Business Week*, etc.).

a) Does the article identify the W's?
b) Discuss whether the display is appropriate for the data.
c) Discuss what the display reveals about the variable and its distribution.
d) Does the article accurately describe and interpret the data? Explain.

24. **Statistics in business, part 2.** Find a graph other than a histogram that shows the distribution of a quantitative variable in a business publication (e.g., *The Wall Street Journal*, *Business Week*, etc.).

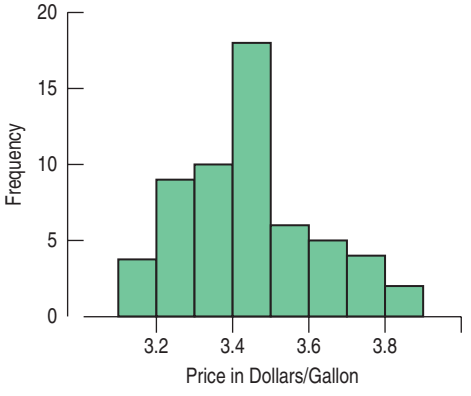
a) Does the article identify the W's?
b) Discuss whether the display is appropriate for the data.
c) Discuss what the display reveals about the variable and its distribution.
d) Does the article accurately describe and interpret the data? Explain.

25. **Two-year college tuition.** The histogram shows the distribution of average tuitions charged by each of the 50 U.S. states for public two-year colleges in the 2007–2008

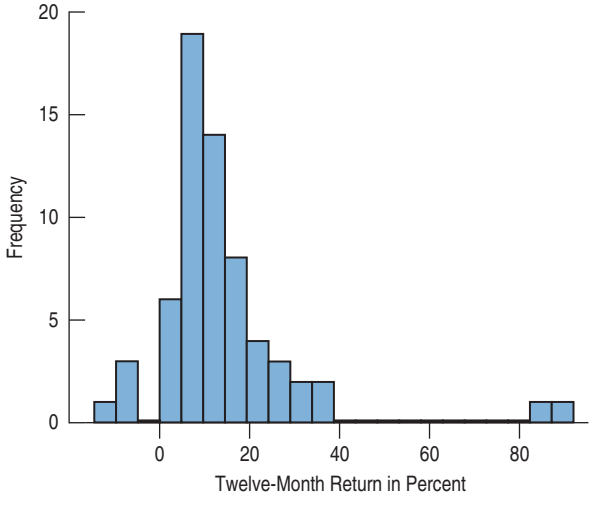


academic year. Write a short description of this distribution (shape, center, spread, unusual features).

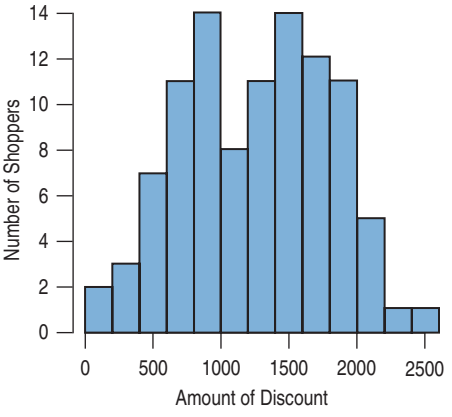
26. Gas prices. The website MSN auto (www.autos.msn.com) provides prices of gasoline at stations all around the United States. This histogram shows the price of regular gas (in $/gallon) for 57 stations in the Los Angeles area during the week before Christmas 2007. Describe the shape of this distribution (shape, center, spread, unusual features).



T **27. Mutual funds.** The histogram displays the 12-month returns (in percent) for a collection of mutual funds in 2007. Give a short summary of this distribution (shape, center, spread, unusual features).



T **28. Car discounts.** A researcher, interested in studying gender differences in negotiations, collects data on the prices that men and women pay for new cars. Here is a histogram of the discounts (the amount in $ below the list price) that men and women received at one car dealership for the last 100 transactions (54 men and 46 women). Give a short summary of this distribution (shape, center, spread, unusual features). What do you think might account for this particular shape?



T **29. Mutual funds, part 2.** Use the data set of Exercise 27 to answer the following questions.

a) Find the five-number summary for these data.
b) Find appropriate measures of center and spread for these data.
c) Create a boxplot for these data.
d) What can you see, if anything, in the histogram that isn't clear in the boxplot?

T **30. Car discounts, part 2.** Use the data set of Exercise 28 to answer the following questions.

a) Find the five-number summary for these data.
b) Create a boxplot for these data.
c) What can you see, if anything, in the histogram of Exercise 28 that isn't clear in the boxplot?

T ***31. Vineyards.** The data set provided contains the data from Exercises 10 and 13. Create a stem-and-leaf display of the sizes of the vineyards in acres. Point out any unusual features of the data that you can see from the stem-and-leaf.

T ***32. Gas prices, again.** The data set provided contains the data from Exercise 26 on the price of gas for 57 stations around Los Angeles in December 2007. Round the data to the nearest penny (e.g., 3.459 becomes 3.46) and create a stem-and-leaf display of the data. Point out any unusual features of the data that you can see from the stem-and-leaf.

33. Gretzky. During his 20 seasons in the National Hockey League, Wayne Gretzky scored 50% more points than anyone else who ever played professional hockey. He accomplished this amazing feat while playing in 280 fewer games than Gordie Howe, the previous record holder. Here are the number of games Gretzky played during each season:

79, 80, 80, 80, 74, 80, 80, 79, 64, 78, 73, 78, 74, 45, 81, 48, 80, 82, 82, 70

*a) Create a stem-and-leaf display.
b) Sketch a boxplot.

c) Briefly describe this distribution.
d) What unusual features do you see in this distribution? What might explain this?

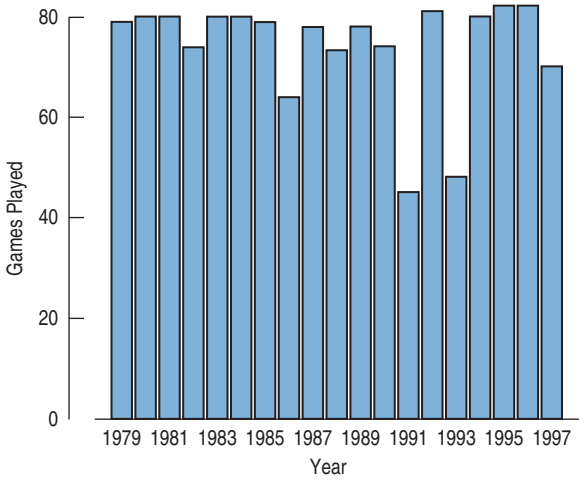
34. McGwire. In his 16-year career as a player in major league baseball, Mark McGwire hit 583 home runs, placing him eighth on the all-time home run list (as of 2008). Here are the number of home runs that McGwire hit for each year from 1986 through 2001:

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

a) *Create a stem-and-leaf display.
b) Sketch a boxplot.
c) Briefly describe this distribution.
d) What unusual features do you see in this distribution? What might explain this?

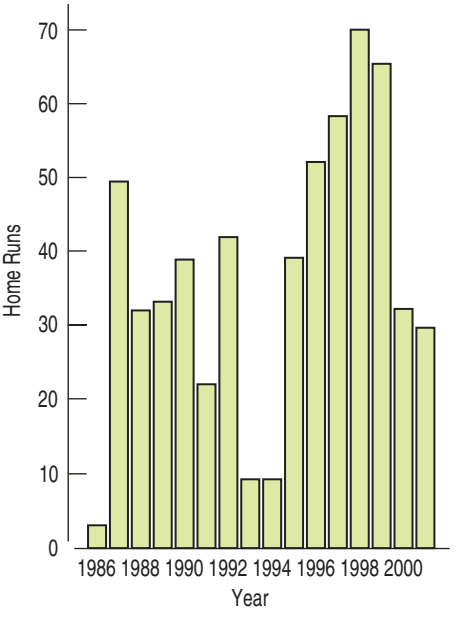
35. Gretzky returns. Look once more at data of hockey games played each season by Wayne Gretzky, seen in Exercise 33.

a) Would you use the mean or the median to summarize the center of this distribution? Why?
b) Without actually finding the mean, would you expect it to be lower or higher than the median? Explain.
c) A student was asked to make a histogram of the data in Exercise 33 and produced the following. Comment.



36. McGwire, again. Look once more at data of home runs hit by Mark McGwire during his 16-year career as seen in Exercise 34.

a) Would you use the mean or the median to summarize the center of this distribution? Why?
b) Find the median.
c) Without actually finding the mean, would you expect it to be lower or higher than the median? Explain.
d) A student was asked to make a histogram of the data in Exercise 34 and produced the following. Comment.



T **37. Pizza prices.** The weekly prices of one brand of frozen pizza over a three-year period in Dallas are provided in the data file. Use the price data to answer the following questions.

a) Find the five-number summary for these data.
b) Find the range and IQR for these data.
c) Create a boxplot for these data.
d) Describe this distribution.
e) Describe any unusual observations.

T **38. Pizza prices, part 2.** The weekly prices of one brand of frozen pizza over a three-year period in Chicago are provided in the data file. Use the price data to answer the following questions.

a) Find the five-number summary for these data.
b) Find the range and IQR for these data.
c) Create a boxplot for these data.
d) Describe the shape (center and spread) of this distribution.
e) Describe any unusual observations.

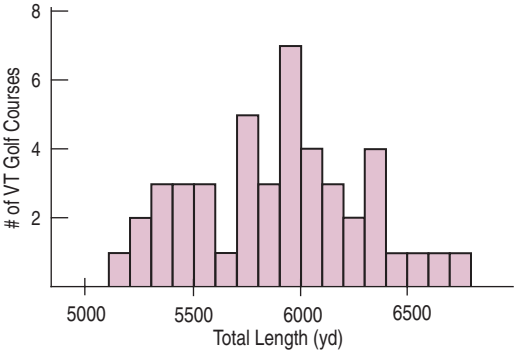
T **39. Gasoline usage.** The U.S. Department of Transportation collects data on the amount of gasoline sold in each state and the District of Columbia. The following data show the per capita (gallons used per person) consumption in the year 2005. Write a report on the gasoline usage by state in the year 2005, being sure to include appropriate graphical displays and summary statistics.

State	Gasoline Usage	State	Gasoline Usage
Alabama	556.91	Montana	486.15
Alaska	398.99	Nebraska	439.46
Arizona	487.52	Nevada	484.26
Arkansas	491.85	New Hampshire	521.45
California	434.11	New Jersey	481.79
Colorado	448.33	New Mexico	482.33
Connecticut	441.39	New York	283.73
Delaware	514.78	North Carolina	491.07
District of Columbia	209.47	North Dakota	513.16
Florida	485.73	Ohio	434.65
Georgia	560.90	Oklahoma	501.12
Hawaii	352.02	Oregon	415.67
Idaho	414.17	Pennsylvania	402.85
Illinois	392.13	Rhode Island	341.67
Indiana	497.35	South Carolina	570.24
Iowa	509.13	South Dakota	498.36
Kansas	399.72	Tennessee	509.77
Kentucky	511.30	Texas	505.39
Louisiana	489.84	Utah	409.93
Maine	531.77	Vermont	537.94
Maryland	471.52	Virginia	518.06
Massachusetts	427.52	Washington	423.32
Michigan	470.89	West Virginia	444.22
Minnesota	504.03	Wisconsin	440.45
Mississippi	539.39	Wyoming	589.18
Missouri	530.72		

40. **OECD** Established in Paris in 1961, the Organisation for Economic Co-operation and Development (OECD) (www.oced.org) collects information on many economic and social aspects of countries around the world. Here are the 2005 gross domestic product (GDP) growth rates (in percentages) of 30 industrialized countries. Write a brief report on the 2005 GDP growth rates of these countries being sure to include appropriate graphical displays and summary statistics.

Country	Growth Rate
Turkey	0.074
Czech Republic	0.061
Slovakia	0.061
Iceland	0.055
Ireland	0.055
Hungary	0.041
Korea, Republic of (South Korea)	0.040
Luxembourg	0.040
Greece	0.037
Poland	0.034
Spain	0.034
Denmark	0.032
United States	0.032
Mexico	0.030
Canada	0.029
Finland	0.029
Sweden	0.027
Japan	0.026
Australia	0.025
New Zealand	0.023
Norway	0.023
Austria	0.020
Switzerland	0.019
United Kingdom	0.019
Belgium	0.015
The Netherlands	0.015
France	0.012
Germany	0.009
Portugal	0.004
Italy	0.000

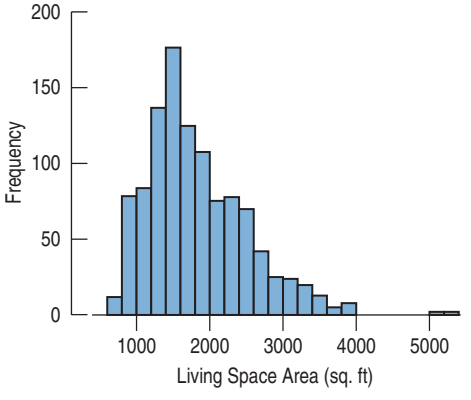
41. **Golf courses.** A start-up company is planning to build a new golf course. For marketing purposes, the company would like to be able to advertise the new course as one of the more difficult courses in the state of Vermont. One measure of the difficulty of a golf course is its length: the total distance (in yards) from tee to hole for all 18 holes. Here are the histogram and summary statistics for the lengths of all the golf courses in Vermont.



Count	45
Mean	5892.91 yd
StdDev	386.59
Min	5185
Q1	5585.75
Median	5928
Q3	6131
Max	6796

a) What is the range of these lengths?
b) Between what lengths do the central 50% of these courses lie?
c) What summary statistics would you use to describe these data?
d) Write a brief description of these data (shape, center, and spread).

42. Real estate. A real estate agent has surveyed houses in 20 nearby zip codes in an attempt to put together a comparison for a new property that she would like to put on the market. She knows that the size of the living area of a house is a strong factor in the price, and she'd like to market this house as being one of the biggest in the area. Here is a histogram and summary statistics for the sizes of all the houses in the area.



Count	1057
Mean	1819.498 sq. ft
Std Dev	662.9414
Min	672
Q1	1342
Median	1675
Q3	2223
Max	5228
Missing	0

a) What is the range of these sizes?
b) Between what sizes do the central 50% of these houses lie?
c) What summary statistics would you use to describe these data?
d) Write a brief description of these data (shape, center, and spread).

T **43. Food sales.** Sales (in $) for one week were collected for 18 stores in a food store chain in the northeastern United States. The stores and the towns they are located in vary in size.

a) Make a suitable display of the sales from the data provided.
b) Summarize the central value for sales for this week with a median and mean. Why do they differ?
c) Given what you know about the distribution, which of these measures does the better job of summarizing the stores' sales? Why?
d) Summarize the spread of the sales distribution with a standard deviation and with an IQR.
e) Given what you know about the distribution, which of these measures does the better job of summarizing the spread of stores' sales? Why?
f) If we were to remove the outliers from the data, how would you expect the mean, median, standard deviation, and IQR to change?

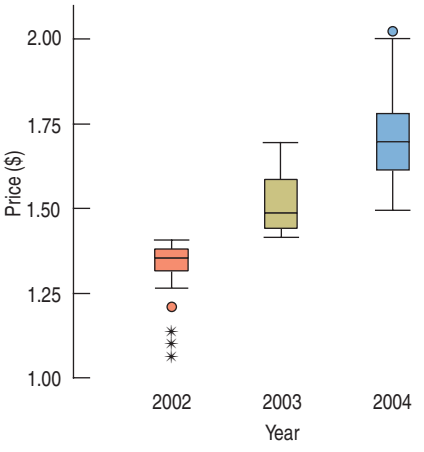
T **44. Insurance profits.** Insurance companies don't know whether a policy they've written is profitable until the policy matures (expires). To see how they've performed recently, an analyst looked at mature policies and investigated the net profit to the company (in $).

a) Make a suitable display of the profits from the data provided.
b) Summarize the central value for the profits with a median and mean. Why do they differ?
c) Given what you know about the distribution, which of these measures might do a better job of summarizing the company's profits? Why?
d) Summarize the spread of the profit distribution with a standard deviation and with an IQR.
e) Given what you know about the distribution, which of these measures might do a better job of summarizing the spread in the company's profits? Why?
f) If we were to remove the outliers from the data, how would you expect the mean, median, standard deviation, and IQR to change?

T **45. iPod failures.** MacInTouch (www.macintouch.com/reliability/ipodfailures.html) surveyed readers about the reliability of their iPods. Of the 8926 iPods owned, 7510 were problem-free while the other 1416 failed. From the data on the CD, compute the failure rate for each of the 17 iPod models. Produce an appropriate graphical display of the failure rates and briefly describe the distribution. (To calculate the failure rate, divide the number failed by the sum of the number failed and the number OK for each model and then multiply by 100.)

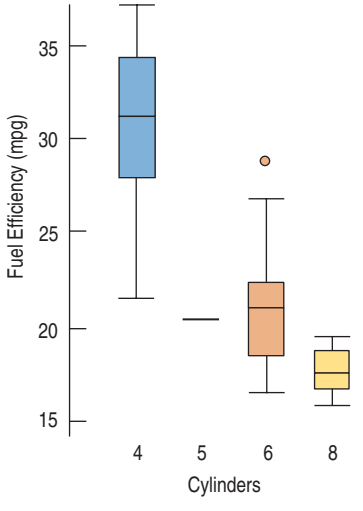
T **46. Unemployment.** The data set provided contains 2008 unemployment rates for 23 developed countries (www.oecd.org). Produce an appropriate graphical display and briefly describe the distribution of unemployment rates.

47. Gas prices, part 2. Below are boxplots of weekly gas prices at a service station in the Midwest United States (prices in $ per gallon).

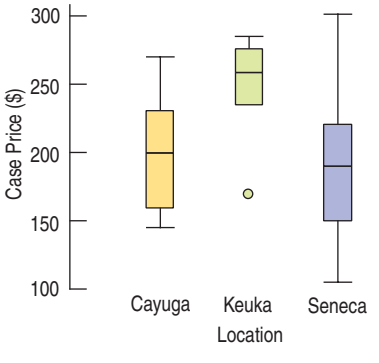


a) Compare the distribution of prices over the three years.
b) In which year were the prices least stable (most volatile)? Explain.

48. Fuel economy. American automobile companies are becoming more motivated to improve the fuel efficiency of the automobiles they produce. It is well known that fuel efficiency is impacted by many characteristics of the car. 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).

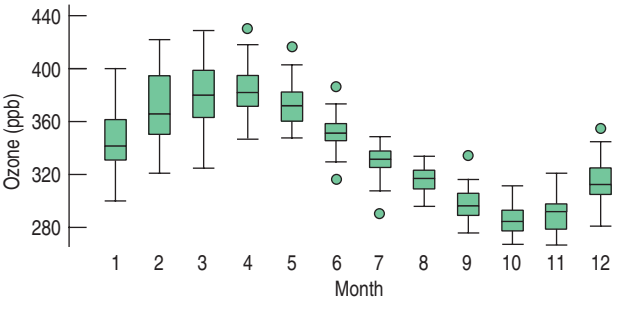


49. Wine prices. The boxplots display case prices (in dollars) of wines produced by vineyards along three of the Finger Lakes in upstate New York.



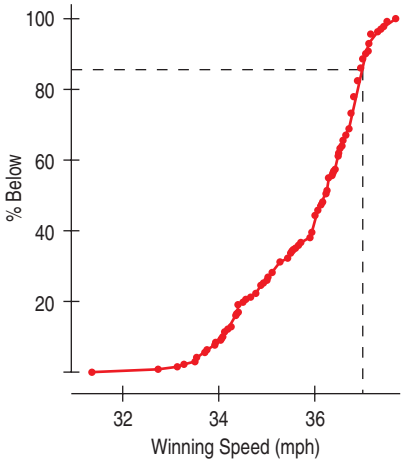
a) Which lake region produces the most expensive wine?
b) Which lake region produces the cheapest wine?
c) In which region are the wines generally more expensive?
d) Write a few sentences describing these prices.

50. 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).



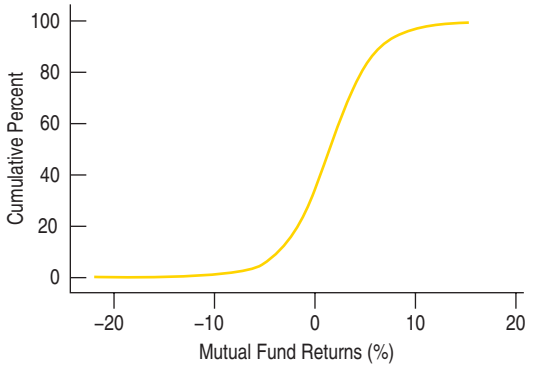
a) In what month was the highest ozone level ever recorded?
b) Which month has the largest IQR?
c) Which month has the smallest range?
d) Write a brief comparison of the ozone levels in January and June.
e) Write a report on the annual patterns you see in the ozone levels.

51. Derby speeds. How fast do horses run? Kentucky Derby winners top 30 miles per hour, as shown in the graph. This graph shows the percentage of Kentucky Derby winners that have run *slower* than a given speed. Note that few have won running less than 33 miles per hour, but about 95% of the winning horses have run less than 37 miles per hour. (A cumulative frequency graph like this is called an **ogive**.)



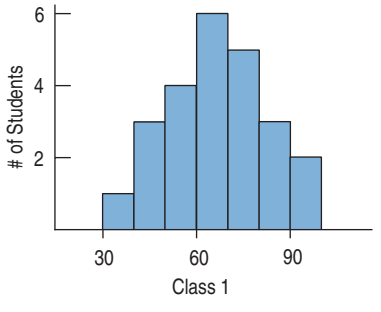
a) Estimate the median winning speed.
b) Estimate the quartiles.
c) Estimate the range and the IQR.
d) Create a boxplot of these speeds.
e) Write a few sentences about the speeds of the Kentucky Derby winners.

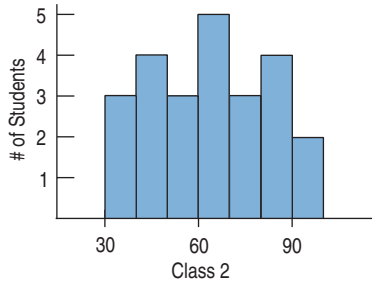
52. Mutual fund, part 3. Here is an ogive of the distribution of monthly returns for a group of aggressive (or high growth) mutual funds over a period of 25 years from 1975 to 1999. (Recall from Exercise 51 that an ogive, or cumulative relative frequency graph, shows the percent of cases at or below a certain value. Thus this graph always begins at 0% and ends at 100%.)

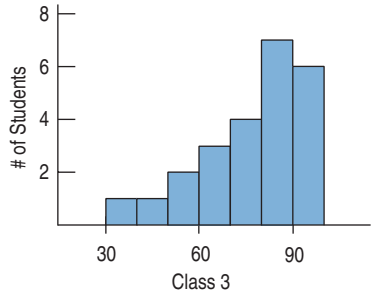


a) Estimate the median.
b) Estimate the quartiles.
c) Estimate the range and the IQR.
d) Create a boxplot of these returns.

53. Test scores. Three Statistics classes all took the same test. Here are histograms of the scores for each class.



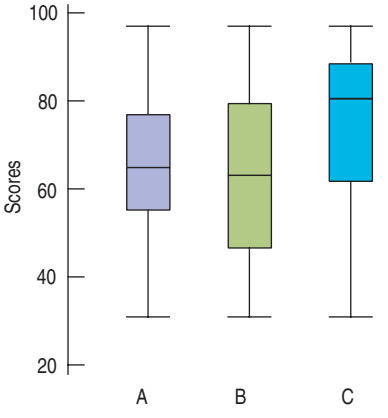




a) Which class had the highest mean score?
b) Which class had the highest median score?
c) For which class are the mean and median most different? Which is higher? Why?
d) Which class had the smallest standard deviation?
e) Which class had the smallest IQR?

54. Test scores, again. Look again at the histograms of test scores for the three Statistics classes in Exercise 53.

a) Overall, which class do you think performed better on the test? Why?
b) How would you describe the shape of each distribution?
c) Match each class with the corresponding boxplot.



T **55. Quality control holes.** 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.

	Fast	Slow
	0.000101	0.000098
	0.000102	0.000096
	0.000100	0.000097
	0.000102	0.000095
	0.000101	0.000094
	0.000103	0.000098
	0.000104	0.000096
	0.000102	0.975600
	0.000102	0.000097
	0.000100	0.000096
Mean	0.000102	0.097647
StdDev	0.000001	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 **56. Fire sale.** A real estate agent notices that houses with fireplaces often fetch a premium in the market and wants to

assess the difference in sales price of 60 homes that recently sold. The data and summary are shown in the table.

	No Fireplace	Fireplace
	142,212	134,865
	206,512	118,007
	50,709	138,297
	108,794	129,470
	68,353	309,808
	123,266	157,946
	80,248	173,723
	135,708	140,510
	122,221	151,917
	128,440	235,105,000
	221,925	259,999
	65,325	211,517
	87,588	102,068
	88,207	115,659
	148,246	145,583
	205,073	116,289
	185,323	238,792
	71,904	310,696
	199,684	139,079
	81,762	109,578
	45,004	89,893
	62,105	132,311
	79,893	131,411
	88,770	158,863
	115,312	130,490
	118,952	178,767
		82,556
		122,221
		84,291
		206,512
		105,363
		103,508
		157,513
		103,861
Mean	**116,597.54**	**7,061,657.74**
Median	**112,053**	**136,581**

Write a report summarizing the findings of the investigation. Include appropriate visual and verbal displays of the distributions, and make a recommendation to the agent about the average premium that a fireplace is worth in this market.

57. **Customer database.** A philanthropic organization has a database of millions of donors that they contact by mail to raise money for charities. One of the variables in the database, *Title*, contains the title of the person or persons printed on the address label. The most common are Mr., Ms., Miss, and Mrs., but there are also Ambassador and Mrs., Your Imperial Majesty, and Cardinal, to name a few others. In all there are over 100 different titles, each with a corresponding numeric code. Here are a few of them.

Code	Title
000	MR.
001	MRS.
1002	MR. and MRS.
003	MISS
004	DR.
005	MADAME
006	SERGEANT
009	RABBI
010	PROFESSOR
126	PRINCE
127	PRINCESS
128	CHIEF
129	BARON
130	SHEIK
131	PRINCE AND PRINCESS
132	YOUR IMPERIAL MAJESTY
135	M. ET MME.
210	PROF.
⋮	⋮

An intern who was asked to analyze the organization's fundraising efforts presented these summary statistics for the variable *Title*.

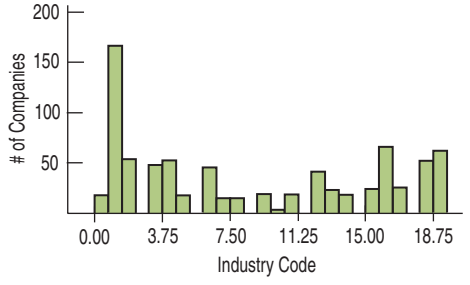
Mean	54.41
StdDev	957.62
Median	1
IQR	2
n	94649

a) What does the mean of 54.41 mean?
b) What are the typical reasons that cause measures of center and spread to be as different as those in this table?
c) Is that why these are so different?

58. **CEOs.** For each CEO, a code is listed that corresponds to the industry of the CEO's company. Here are a few of the codes and the industries to which they correspond.

Industry	Industry Code	Industry	Industry Code
Financial services	1	Energy	12
Food/drink/tobacco	2	Capital goods	14
Health	3	Computers/ communications	16
Insurance	4	Entertainment/ information	17
Retailing	6	Consumer non-durables	18
Forest products	9	Electric utilities	19
Aerospace/defense	11		

A recently hired investment analyst has been assigned to examine the industries and the compensations of the CEOs. To start the analysis, he produces the following histogram of industry codes.



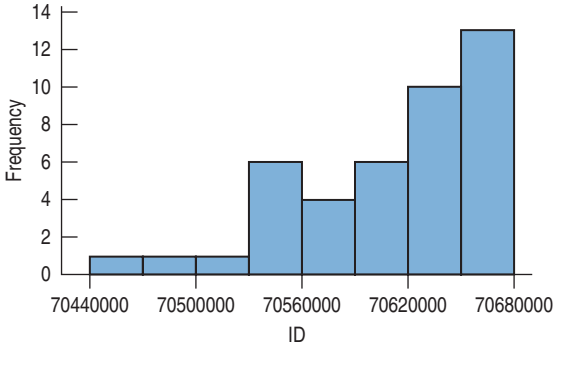
a) What might account for the gaps seen in the histogram?
b) What advice might you give the analyst about the appropriateness of this display?

T **59. Mutual funds types.** The 64 mutual funds of Exercise 27 are classified into three types: U.S. Domestic Large Cap Funds, U.S. Domestic Small/Mid Cap Funds, and International Funds. Compare the 3-month return of the three types of funds using an appropriate display and write a brief summary of the differences.

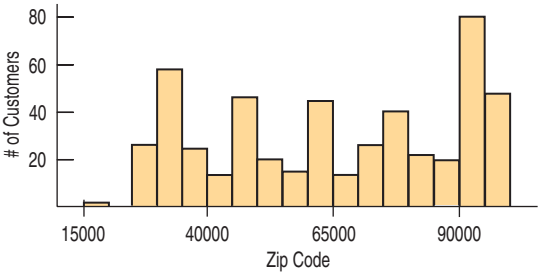
T **60. Car discounts, part 3.** The discounts negotiated by the car buyers in Exercise 28 are classified by whether the buyer was Male (code = 0) or Female (code = 1). Compare the discounts of men vs. women using an appropriate display and write a brief summary of the differences.

61. Houses for sale. Each house listed on the multiple listing service (MLS) is assigned a sequential ID number. A recently hired real estate agent decided to examine the MLS numbers in a recent random sample of homes for sale by one real estate agency in nearby towns. To begin the analysis, the agent produces the following histogram of ID numbers.

a) What might account for the distribution seen in the histogram?
b) What advice might you give the analyst about the appropriateness of this display?



62. Zip codes. Holes-R-Us, an Internet company that sells piercing jewelry, keeps transaction records on its sales. At a recent sales meeting, one of the staff presented the following histogram and summary statistics of the zip codes of the last 500 customers, so that the staff might understand where sales are coming from. Comment on the usefulness and appropriateness of this display.



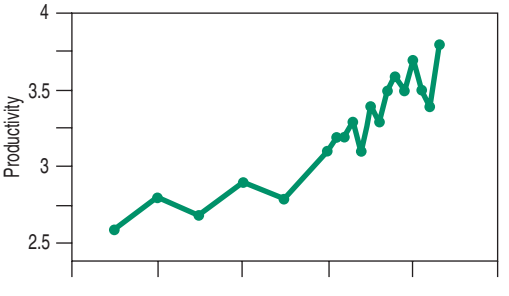
T ***63. Hurricanes.** Buying insurance for property loss from hurricanes has become increasingly difficult since hurricane Katrina caused record property loss damage. Many companies have refused to renew policies or write new ones. The data set provided contains the total number of hurricanes by every full decade from 1851 to 2000 (from the National Hurricane Center). Some scientists claim that there has been an increase in the number of hurricanes in recent years.

a) Create a histogram of these data.
b) Describe the distribution.
c) Create a time series plot of these data.
d) Discuss the time series plot. Does this graph support the claim of these scientists, at least up to the year 2000?

T ***64. Hurricanes, part 2.** Using the hurricanes data set, examine the number of major hurricanes (category 3, 4, or 5) by every full decade from 1851 to 2000.

a) Create a histogram of these data.
b) Describe the distribution.
c) Create a timeplot of these data.
d) Discuss the timeplot. Does this graph support the claim of scientists that the number of major hurricanes has been increasing (at least up through the year 2000)?

65. Productivity study. The National Center for Productivity releases information on the efficiency of workers. In a recent report, they included the following graph showing a rapid rise in productivity. What questions do you have about this?



66. Productivity study revisited. A second report by the National Center for Productivity analyzed the relationship between productivity and wages. Comment on the graph they used.

67. Real estate, part 2. The 1057 houses described in Exercise 42 have a mean price of $167,900, with a standard deviation of $77,158. The mean living area is 1819 sq. ft., with a standard deviation of 663 sq. ft. Which is more unusual, a house in that market that sells for $400,000 or a house that has 4000 sq. ft of living area? Explain.

68. Tuition, 2008. The data set provided contains the average tuition of private four-year colleges and universities as well as the average 2007–2008 tuitions for each state seen in Exercise 25. The mean tuition charged by a public two-year college was $2763, with a standard deviation of $988. For private four-year colleges the mean was $21,259, with a standard deviation of $6241. Which would be more unusual: a state whose average public two-year college is $700 or a state whose average private four-year college tuition was $10,000? Explain.

69. Food consumption. FAOSTAT, the Food and Agriculture Organization of the United Nations, collects information on the production and consumption of more than 200 food and agricultural products for 200 countries around the world. Here are two tables, one for meat consumption (per capita in kg per year) and one for alcohol consumption (per capita in gallons per year). The United States leads in meat consumption with 267.30 pounds, while Ireland is the largest alcohol consumer at 55.80 gallons.

Country	Alcohol	Meat	Country	Alcohol	Meat
Australia	29.56	242.22	Luxembourg	34.32	197.34
Austria	40.46	242.22	Mexico	13.52	126.50
Belgium	34.32	197.34	Netherlands	23.87	201.08
Canada	26.62	219.56	New Zealand	25.22	228.58
Czech Republic	43.81	166.98	Norway	17.58	129.80
Denmark	40.59	256.96	Poland	20.70	155.10
Finland	25.01	146.08	Portugal	33.02	194.92
France	24.88	225.28	Slovakia	26.49	121.88
Germany	37.44	182.82	South Korea	17.60	93.06
Greece	17.68	201.30	Spain	28.05	259.82
Hungary	29.25	179.52	Sweden	20.07	155.32
Iceland	15.94	178.20	Switzerland	25.32	159.72
Ireland	55.80	194.26	Turkey	3.28	42.68
Italy	21.68	200.64	United Kingdom	30.32	171.16
Japan	14.59	93.28	United States	26.36	267.30

Using z-scores, find which country is the larger consumer of both meat and alcohol together.

70. World Bank. The World Bank, through their Doing Business project (www.doingbusiness.org), ranks nearly 200 economies on the ease of doing business. One of their rankings measures the ease of starting a business and is made up (in part) of the following variables: number of required start-up procedures, average start-up time (in days), and average start-up cost (in % of per capita income). The following table gives the mean and standard deviations of these variables for 95 economies.

	Procedures (#)	Time (Days)	Cost (%)
Mean	7.9	27.9	14.2
SD	2.9	19.6	12.9

Here are the data for three countries.

	Procedures	Time	Cost
Spain	10	47	15.1
Guatemala	11	26	47.3
Fiji	8	46	25.3

a) Use z-scores to combine the three measures.
b) Which country has the best environment after combining the three measures? Be careful—a lower rank indicates a better environment to start up a business.

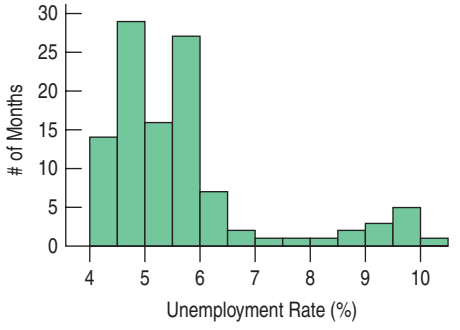
***71. Regular gas.** The data set provided contains U.S. regular retail gasoline prices (cents/gallon) from August 20, 1990 to May 28, 2007, from a national sample of gasoline stations obtained from the U.S. Department of Energy.

a) Create a histogram of the data and describe the distribution.
b) Create a time series plot of the data and describe the trend.
c) Which graphical display seems the more appropriate for these data? Explain.

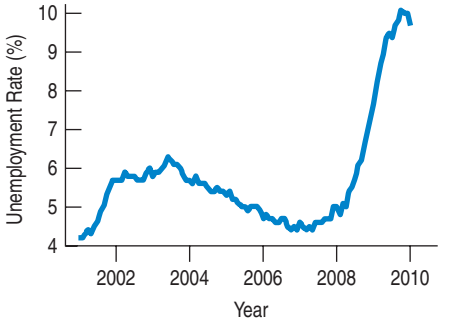
***72. Home price index.** Standard and Poor's Case-Shiller® Home Price Index measures the residential housing market in 20 metropolitan regions across the United States. The national index is a composite of the 20 regions and can be found in the data set provided.

a) Create a histogram of the data and describe the distribution.
b) Create a time series plot of the data and describe the trend.
c) Which graphical display seems the more appropriate for these data? Explain.

***73. Unemployment rate, 2010.** The histogram shows the monthly U.S. unemployment rate from January 2001 to January 2010.

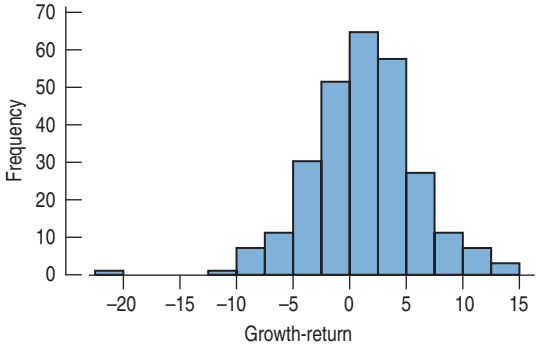


Here is the time series plot for the same data.

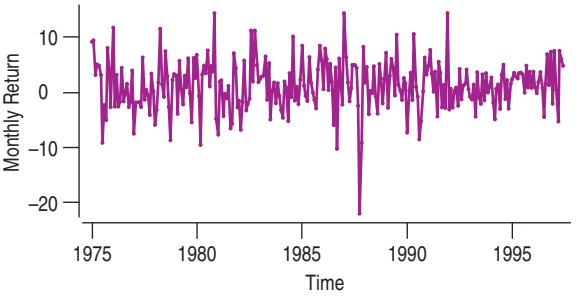


a) What features of the data can you see in the histogram that aren't clear in the time series plot?
b) What features of the data can you see in the time series plot that aren't clear in the histogram?
c) Which graphical display seems the more appropriate for these data? Explain.
d) Write a brief description of unemployment rates over this time period in the United States.

***74. Mutual fund performance.** The following histogram displays the monthly returns for a group of mutual funds considered aggressive (or high growth) over a period of 22 years from 1975 to 1997.

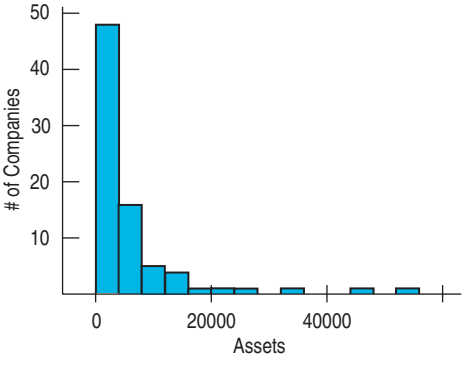


Here is the time series plot for the same data.



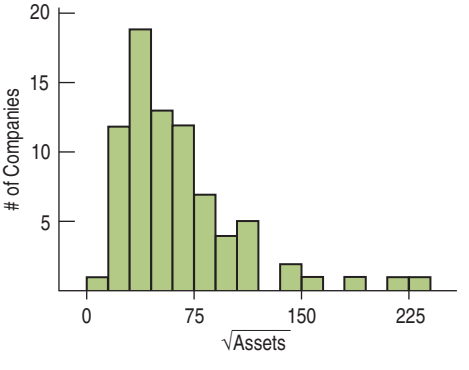
a) What features of the data can you see in the histogram that aren't clear from the time series plot?
b) What features of the data can you see in the time series plot that aren't clear in the histogram?
c) Which graphical display seems the more appropriate for these data? Explain.
d) Write a brief description of monthly returns over this time period.

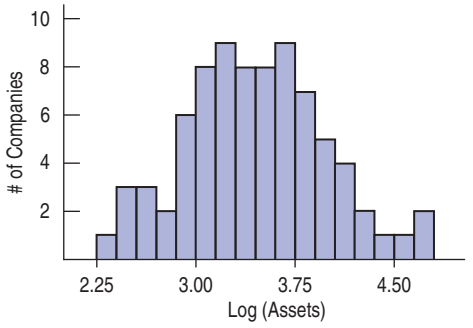
75. Assets. Here is a histrogram of the assets (in millions of dollars) of 79 companies chosen from the *Forbes* list of the nation's top corporations.



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?

76. Assets, again. Here are the same data you saw in Exercise 75 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?

Just Checking Answers

1. Incomes are probably skewed to the right and not symmetric, making the median the more appropriate measure of center. The mean will be influenced by the high end of family incomes and not reflect the "typical" family income as well as the median would. It will give the impression that the typical income is higher than it is.
2. An IQR of 30 mpg would mean that only 50% of the cars get gas mileages in an interval 30 mpg wide. Fuel economy doesn't vary that much. 3 mpg is reasonable. It seems plausible that 50% of the cars will be within about 3 mpg of each other. An IQR of 0.3 mpg would mean that the gas mileage of half the cars varies little from the estimate. It's unlikely that cars, drivers, and driving conditions are that consistent.
3. We'd prefer a standard deviation of 2 months. Making a consistent product is important for quality. Customers want to be able to count on the MP3 player lasting somewhere close to 5 years, and a standard deviation of 2 years would mean that life spans were highly variable.