5

Displaying and Describing Quantitative Data

LEARNING OBJECTIVES

In this chapter we show you how to display quantitative data graphically and how to analyze that display. After reading and studying this chapter, you should be able to:

1. Display data in a histogram and in a stem-and-leaf diagram
2. Estimate the "centre" of the data distribution
3. Estimate the spread of the data distribution
4. Graph the centre of the data distribution and the extent to which it's spread in a "boxplot"
5. Identify outliers
6. Standardize data relative to its spread
7. Graph time series data

Bell Canada

Alexander Graham Bell, who was born in Scotland in 1847, is well known today as the inventor of the telephone. (He wasn't always known by this name, though: His two brothers had been given middle names, and Alexander wanted one too. For his 11th birthday present his parents gave him the middle name Graham, and the rest is history.) He moved to Canada at the age of 23 and worked partly in Brantford, Ontario, and partly in Boston, Massachusetts, where he raced Eliza Gray to the patent office and received patent #174,465 for the telephone.

Alexander licensed the patent to various companies, including the Bell Telephone Company of Canada, which in 1880 was given a monopoly to provide long distance service in Canada. The telephone

equipment, initially manufactured in-house, began to be manufactured in a spinoff company named Northern Electric in 1896. These two companies eventually formed Bell Canada and Nortel Networks, with the former purchasing equipment from the latter.

However, this close relationship ended in the 1990s, when Nortel's price for ATM (Asynchronous Transfer Mode) switches was far above its competitors' price; Bell Canada bought its equipment from General DataComm (GDC). Nortel ultimately went bankrupt in 2009, but Bell Canada continues to thrive, with over 50,000 employees and over $18 billion in revenues in 2010.

To learn more about the behaviour of Bell Canada's stock, let's start by looking at Table 5.1, which gives the daily changes in stock price (in Canadian dollars) over a two-month period.

Daily Changes in Bell Canada Stock Price ($) September 12–30, 2011	October 3–21, 2011
+0.51	+0.06
+0.1	–0.07
–0.21	–0.03
0	+0.64
+0.29	–0.36
+0.4	+0.02
–0.12	–0.42
–0.33	+0.59
+0.37	+0.39
–0.77	+0.19
–0.03	–0.17
0.19	0
+0.49	–0.33
–0.36	–0.19
–0.15	

Table 5.1 Daily price changes in Bell Canada stock for the period September 12 to October 21, 2011.

It's hard to tell very much from tables of values like this. We might get a rough idea of how much the stock changed from day to day—usually less than $0.40 in either direction—but that's about it. In what other way might we display this data?

L.O. ❶

5.1 Displaying Data Distributions

WHO	Days
WHAT	Daily changes in Bell Canada's stock price in dollars
WHEN	September 12 to October 21, 2011
WHERE	Toronto Stock Exchange
WHY	To examine Bell Canada stock volatility

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 Bell's stock price change is a *quantitative* variable, whose units are dollars.

Histograms

Figure 5.1 shows the daily price changes of Bell Canada stock displayed as a frequency distribution and a histogram.

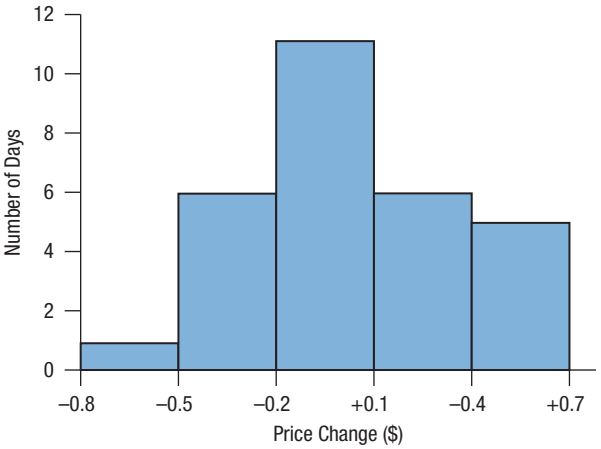
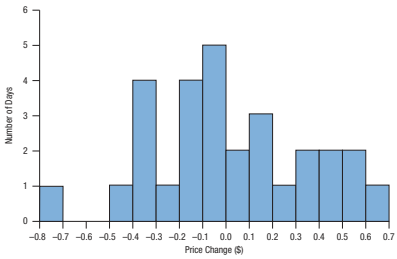
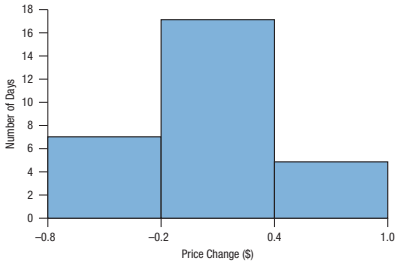


Figure 5.1 Daily price changes of Bell Canada stock. The histogram displays the distribution of price changes by showing, for each "bin" of price changes, the number of days having price changes in that bin.

Like a bar chart, a **histogram** plots the bin counts as the heights of bars. In this histogram of daily price changes, each bin has a width of $0.30, so, for example, the height of the tallest bar says that there were 11 daily price changes of between −$0.20 and +$0.10. In this way, the histogram displays the entire distribution of price changes. Unlike a bar chart, which puts gaps between bars to separate the categories, no gaps appear between the bars of a histogram *unless* there are actual gaps in the data. Gaps can be important, so watch out for them.

For categorical variables, each category is represented by its own bar. That was easy; there was no choice, except maybe to combine categories for ease of display. But for quantitative variables, we have to choose how to slice up all the possible values into bins. Once we have equal-width bins, the histogram can count the number of cases that fall into each bin, represent the counts as bars, and plot them against the bin values. In this way, it displays the distribution at a glance.





- **How do histograms work?** If you were to make a histogram by hand or in Excel, you'd need to make some decisions about the bins. First, you would need to decide how wide to make the bins. The width of bins is important, because some features of the distribution may appear more obvious at different bin width choices. One rule of thumb is that the number of bins depends on how much data we have. If we have n data points, we use about $\log_2 n$ bins.
  In our case, with $n = 29$ data points, $\log_2 n = 4.86$, so we have rounded off to five and used five bins. This is not a rule that's cast in stone. More bins will give more detail. Fewer bins will give a smoother histogram. It's your choice.

  However, if we use *too many* bins (as in the upper graph on the left with 15 bins), the histogram will look pretty random and the overall shape of Figure 5.1 will be lost. With too few bins (three bins in the lower graph on the left), we lose a lot of information. For example, there are not in fact any days with price changes between $0.70 and $1.00, even though we can't tell that from the histogram.

  With many statistics packages, you can easily vary the bin width interactively, so that you can make sure that a feature you think you see isn't just a consequence of a certain choice of bin width.

  Next you'd need to decide where to place the endpoints of the bins. (Statistics packages and graphing calculators make these choices for you automatically.) Bins are always equal in width. But what do you do with a value of $5 if one bin spans from $0 to $5 and the next bin spans from $5 to $10? It's important to have a consistent rule for a value that falls exactly on a bin boundary; so, for example, you'd put a month with a change of $5 into the $5 to $10 bin rather than the $0 to $5 bin. That said, the purpose of a histogram is to describe the overall "shape" of our data, not to worry too much about individual data values.

From the histogram in Figure 5.1, we can see that the daily price changes were around $0.00. We can also see that, although they vary, most of the daily price changes were between –$0.50 and +$0.70. On one day only was the change less than –$0.50.

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 (see Figure 5.2). The shape of the histogram is exactly the same (as in Figure 5.1); only the labels are different.

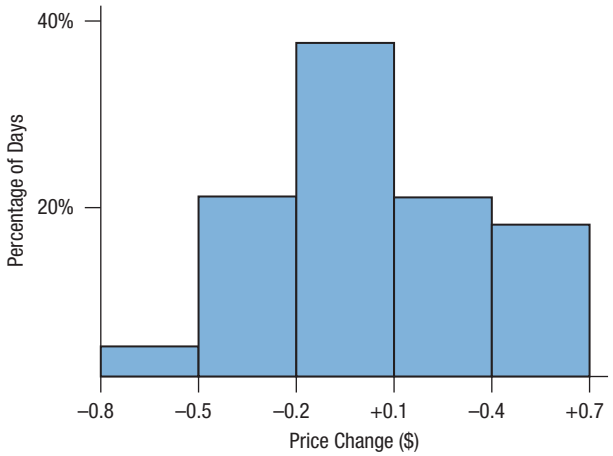


Figure 5.2 A relative frequency histogram looks just like a frequency histogram except that the vertical axis now shows the percentage of months in each bin.

For Example Creating a histogram

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:

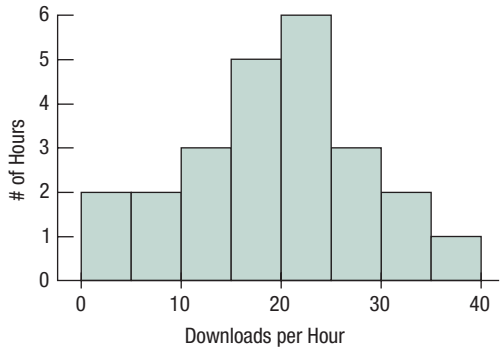
HOUR	DOWNLOADS PER HOUR	HOUR	DOWNLOADS PER HOUR
12:00 A.M.	36	12:00 P.M.	25
1:00 A.M.	28	1:00 P.M.	22
2:00 A.M.	19	2:00 P.M.	17
3:00 A.M.	10	3:00 P.M.	18
4:00 A.M.	5	4:00 P.M.	20
5:00 A.M.	3	5:00 P.M.	23
6:00 A.M.	2	6:00 P.M.	21
7:00 A.M.	6	7:00 P.M.	18
8:00 A.M.	12	8:00 P.M.	24
9:00 A.M.	14	9:00 P.M.	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: There are 24 data points, and $\text{Log}_2\ 24 = 4.6$, so we need about four or five bins. The data are in the 0 to 40 range, so it makes sense to use four bins of width 10. The easiest way to do this is to first put the data in order: 2, 3, 5, 6, 10, 12, 14, 17, 18, 18, 18, 19, 20, 20, 21, 22, 23, 24, 25, 27, 28, 30, 30, 36, and then make a frequency table:

DOWNLOADS PER HOUR	NUMBER OF HOURS
0–10	4
10–20	8
20–30	9
30–40	3
Total	**24**

The histogram will look 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] Figure 5.3 compares a stem-and-leaf display with a histogram for some other data on stock price changes. We've used more bins than we normally would in order to provide a detailed example with a small amount of data. As you can see, a stem-and-leaf display is basically a histogram turned on its side.

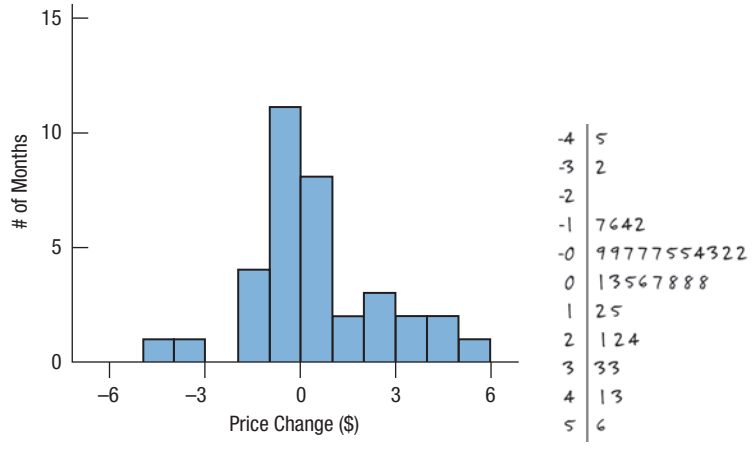


Figure 5.3 Thirty-six months of monthly stock price changes displayed by both a histogram (left) and a stem-and-leaf display (right).

[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 that emerge from just one quick look at the distribution of a quantitative variable are often quite valuable.

- **How do stem-and-leaf displays work?** Stem-and-leaf displays use part of each number (called the stem) to name the bins. To make the "leaves," stem-and-leaf diagrams use the next digit of the number. For example, if we had a monthly price change of $2.1, we could write 2|1, where 2 serves as the stem and 1 as the leaf. To display the changes 2.06, 2.22, 2.44, 3.28, and 3.34 together, we would write

2 124

3 33

Notice that we've rounded off the data—for example, 2.06 becomes 2.1, so that only one significant figure is used in the "leaf." Often we put the higher numbers on top, but either way is common. Featuring higher numbers on top is often natural, but putting the higher numbers on the bottom keeps the direction of the histogram the same when you tilt your head to look at it—otherwise, the histogram appears reversed.

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

There are both positive and negative values in the price changes. Values of $0.3 and $0.5 are displayed as leaves of "3" and "5" on the "0" stem. But values of –$0.3 and –$0.5 must be plotted below zero. So the stem-and-leaf display has a "–0" stem to hold them—again with leaves of "3" and "5." It may seem a little strange to see two zero stems, one labelled "20." But if you think about it, you'll see that it's a sensible way to deal with negative values.

Stem-and-leaf displays are great pencil-and-paper constructions and are well suited to moderate amounts of data—say, between 10 and a few hundred values. They retain all the quantitative values that are summarized in the graphics of a histogram, but for larger data sets, histograms do a better job. If you're making a stem-and-leaf diagram from more than 100 data points, you may need to "split" the leaves. In the example above,

01 1356788

could become:

0 135

0 6788

In Chapter 4, you learned to check the Categorical Data Condition. Now, by contrast, before making a stem-and-leaf display or a histogram, you need to check the Quantitative Data Condition: that the data represent values of a quantitative variable.

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.

L.O. ❷

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 centre, and its spread.

We describe the **shape** of a distribution in terms of its mode(s), its symmetry, and whether it has any gaps or outlying values.

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 Bell Canada data? The value \$0.00 occurred twice, but so did –\$0.33. Which of these is the mode? Probably neither. 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.

Does the histogram have a single 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 centre of this hump. The Bell Canada stock price changes have a single mode at just below \$0 (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 Bell Canada price changes, is called **unimodal**; distributions whose histograms have two humps are **bimodal**, and those with three or more are called **multimodal**. For example, Figure 5.4 represents a bimodal distribution.

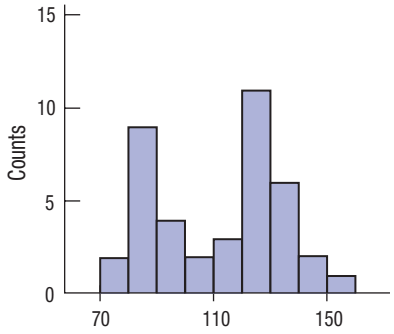


Figure 5.4 A bimodal distribution has two apparent modes.

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 data distribution whose histogram doesn't appear to have any clear mode and in which all the bars are approximately the same height is approximately **uniform** (see Figure 5.5). (Chapter 9 gives a more formal definition.)

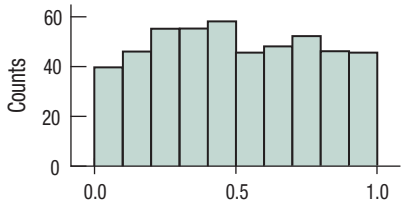


Figure 5.5 In an approximately uniform distribution, bars are all about the same height. The histogram doesn't appear to have a mode.

Symmetry

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

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 data distribution is approximately **symmetric** if it can be divided into two parts that look, at least approximately, like mirror images.

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.

[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.

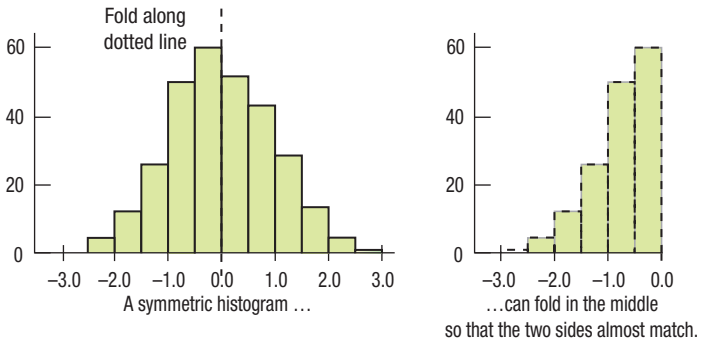


Figure 5.6 An approximately symmetric histogram can be folded in the middle so that the two sides almost match.

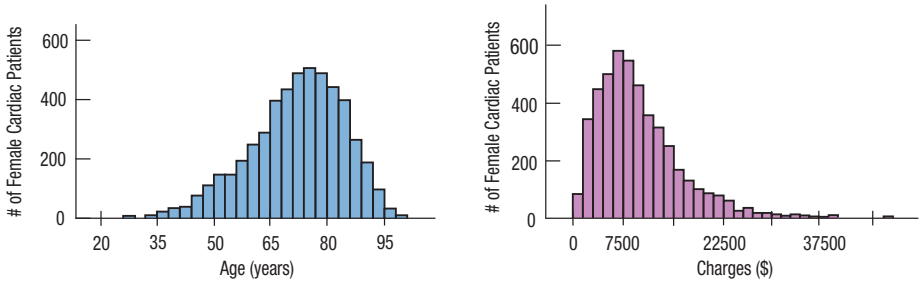


Figure 5.7 Two skewed histograms showing the age (on left) and hospital charges (on right) for all female heart attack patients in New York State in one year. The histogram of *Age* (in blue) is skewed to the left, while the histogram of *Charges* (in purple) is skewed to the right.

Outliers

Do any features appear to stand 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 data distribution. For example, if you're studying personal wealth and Bill Gates is in your sample, he would 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 statistical 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.) In Section 5.8 we'll offer you a rule of thumb for how to decide if and when a value might be considered to be an outlier, along with some advice for what to do when you encounter one.

- **Using your judgment.** How you characterize a data 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 see how persistent some of the features are. 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. Make 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 previous example on page xx.

Answer: It is fairly symmetric and unimodal with no outliers.

L.O. ❷

5.3 Centre

Notation Alert!

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

Look again at the Bell Canada price changes in Figure 5.1. If you had to pick one number to describe a *typical* price change, what would you pick? When a histogram is unimodal and symmetric, most people would point to the **centre** of the distribution, where the histogram peaks. The typical price change is between −$0.20 and +$0.10.

If we want to be more precise and *calculate* a number, we can *average* the data. In the Bell Canada example, the average price change is $0.024, 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 a generic variable y, and use the Greek capital letter sigma, Σ, to mean "sum" (sigma is "S" in Greek), and write[3]

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

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 this value the **mean** of y.[4]

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. For example, Figure 5.7 showed a histogram of the total charges for hospital stays of female heart attack patients in one year in New York State. The mean value is $10,260.70. Locate that value on the histogram. Does it seem a little high as a summary of a typical cost? In fact, about two-thirds of the charges are lower 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. So the median is resistant; it isn't

[3]You may also see the variable called x and the equation written as $\bar{x} = \frac{Total}{n} = \frac{\Sigma x}{n}$. 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 the word *average* is often used too colloquially, as in the "average"homebuyer, where we don't sum up anything. Even though average *is* sometimes used in the way we intend, as in a batting average, we'll often use the more precise term *mean* throughout the book.

affected by unusual observations or by the shape of the distribution. Because of its resistance to these effects, the median is commonly used for variables such as cost or income, which are likely to be skewed. For the female heart attack patient charges, the median cost is $8619, which seems like a more appropriate summary (see Figure 5.8).

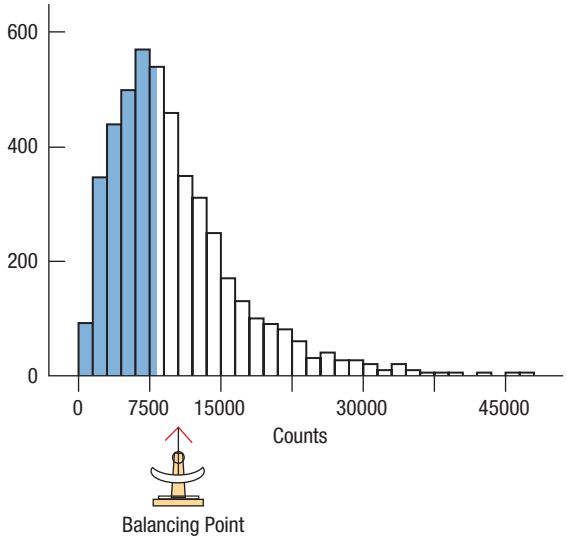


Figure 5.8 The median splits the area of the histogram in half at $8619. Because the distribution is skewed to the right, the mean $10,260 is higher than the median. The points at the right in the tail of the data distribution have pulled the mean toward them, away from the median.

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. Since there are seven values, 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. Now we have 8 values, an even number. The median is the average of the 8/2, or 4th, and the $(8/2) + 1$, or 5th, values. So the median is $(13.9 + 14.1)/2 = 14.0$.

Does it really make a difference whether we choose a mean or a median? The mean price change for the Bell Canada stock is $0.024. Because the distribution

of the price changes is roughly symmetric, we'd expect the mean and median to be close. In fact, we compute the median to be \$0.00. But for variables with skewed distributions, the story is quite different. For a right-skewed distribution like the hospital charges in Figure 5.8, the mean is larger than the median: \$10,260 compared with \$8619. The difference is due to the overall shape of the distributions.

The mean is the point at which the histogram would balance. Just like a child who moves away from the centre of a see-saw, a bar in a histogram that is located far from the centre has more leverage, pulling the mean in its direction. It's hard to argue that the mean, which has been pulled aside by only a few outlying values or by a long tail, is what we mean by the centre 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 to draw more interesting conclusions. It will be the summary value we work with much more throughout the rest of the book.

Geometric Mean

Although the mean is a natural measure of the average of a set of numbers, there are some circumstances in which it would be inappropriate. Suppose you put \$1000 into an investment that grows 10% in the first year, 20% in the second year, and 60% in the third year. The average rate of growth of your investment is *not* $(10 + 20 + 60)/3 = 30$. We can see this by calculating the value of your investment at the end of each of those three years.

End of Year	Growth Rate	Value (\$)
		1000.00
1	10%	1100.00
2	20%	1320.00
3	60%	2112.00

At 30% growth each year, you would have:

End of Year	Growth Rate	Value (\$)
		1000.00
1	30%	1300.00
2	30%	1690.00
3	30%	2197.00

The average rate of growth is the growth rate that generates \$2112 after three years. Suppose this rate is r. Then at the end of year 1, you would have $\$1000 \times (1 + r)$. At the end of year 3, you would have $\$1000 \times (1 + r)^3$. Thus:

$$(1 + 0.1) \times (1 + 0.2) \times (1 + 0.6) = (1 + r)^3$$

or

$$(1 + r) = [(1.1) \times (1.2) \times (1.6)]^{1/3}$$

$$r = 28.3\%$$

We also call this the Compound Annual Growth Rate (CAGR) of your investment.

End of Year	Growth Rate	Value ($)
		1000.00
1	28.30%	1283.01
2	28.30%	1646.12
3	28.30%	2112.00

1.283 is the geometric mean of 1.1, 1.2, and 1.6.

In general, we find the **geometric mean** of a set of n numbers $a_1, a_2, \ldots, a_n$ by multiplying them together and taking the n^{th} root of the product.

$$\text{Geometric Mean} = (a_1 \times a_2 \times \cdots \times a_n)^{1/n}$$

For comparison, the regular mean (sometimes called the arithmetic mean) = $(a_1 \times a_2 \times \cdots \times a_n)^{1/n}$.

For Example Finding the mean and median

Question: From the data in the first example on page xx, what is a typical number of downloads per hour?

Answer: The mean number is 18.7 downloads per hour. Since there are 24 data values, the median is the average of the 12th and 13th values: (19 + 20)/2 = 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 median, but either one seems like a reasonable summary to report.

L.O. ❸

5.4 Spread

We know that the typical price change of Bell Canada stock is around $0, but knowing the mean or median alone doesn't tell us about the entire distribution. A stock whose price change doesn't move away from $0 isn't very interesting. The more the data vary, the less a measure of centre 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 Bell Canada data, the range is $0.64 − (−$0.77) = $1.41. 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 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:

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

We first sort the Bell Canada data from smallest to largest values and get the following figures:

−0.77; −0.42; −0.36; −0.36; −0.33; −0.33; −0.21; −0.19; −0.17; −0.15; −0.12; −0.07; −0.03; −0.03; 0; 0; 0.02; 0.06; 0.1; 0.19; 0.19; 0.29; 0.37; 0.39; 0.4; 0.49; 0.51; 0.59; 0.64

There are 29 values in total. Since 29 is an odd number, we can take the 15th data value as the median, so that there are 14 data values on either side of it. The median is therefore 0. In order to calculate the quartiles, we take the two halves of the data including the median in both halves. The first quartile, Q1, is the median of the lower half (1st to 15th data values), meaning the 8th data value, so that there are 7 data values on either side of it. Q1 therefore equals –0.19. Q3 is the median of data values 15 to 29 (i.e., the 22nd data value which gives Q3 = 0.29). So the IQR = Q3 − Q1 = \$0.29 − (−\$0.19) = \$0.48.

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 approximately 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 so that we don't get any negative values. The average of the *squared* deviations is called the **variance** and is denoted by s^2:

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

The further the individual data values, y, are from the mean, $\bar{y}$, the larger the variance.

Quartiles

An easy way to find the quartiles is to first 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 and use them as the quartiles.

Why do banks favour the formation of a single customer line that feeds several teller windows rather than separate lines for each teller? The average waiting time is less variable when a single line is formed, and people prefer consistency.

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.

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: $s = \sqrt{13.33} = 3.65$.

You may be surprised that we divide by $n - 1$ in this calculation, whereas when we calculated the mean we divided by n. We calculate the variance by dividing by $n - 1$ whenever our data is just a sample of the complete population of data that could potentially be collected. This is usually the case. Our data on the Bell Canada stock price covers only certain days. There's no point in going back into ancient history and collecting stock prices from the day the company was founded, so a recent sample of stock prices is a realistic sample to work with.

The most common situation in which we have complete data on a population is when we're using census data. In that case, the variance is calculated by dividing by n instead of $n - 1$. We use Greek letters for populations: μ for mean and σ for standard deviation.

$$\sigma^2 = \frac{\sum(y - \mu)^2}{n}$$

The above formula assumes that we've already calculated the mean of our data. An equivalent formula that's easier to use when we don't know the mean is

$$s^2 = \frac{y^2 - (\ y)^2/n}{n - 1} \text{ for a sample}$$

or

$$\sigma^2 = \frac{y^2 - (\ y)^2/n}{n} \text{ for a population.}$$

The variance plays an important role in statistics, but as a measure of spread, it's problematic. Whatever the units of the original data, the variance is in *squared* units. We often 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**.

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

For the Bell Canada stock price changes, $s = \$0.34$. We now have measures of centre and spread that are suited to different types of data, as summarized in the following table:

	Centre	Spread
Approximately Symmetric Data	Mean	Standard deviation
Asymmetric Data	Median	Interquartile range

For Example — Describing the spread

Questions: For the data from the first example on page xx, describe the spread of the number of downloads per hour.

Answer: The range of downloads is $36 - 2 = 34$ downloads per hour.

The first quartile, Q1, is the median of the first 12 data points (i.e., the average of the 6th and 7th): $Q1 = (12 + 14)/2 = 13$. Linewise, $Q3 = (24 + 25)/2 = 24.5$. So the IQR is $24.5 - 13 = 11.5$ downloads per hour. The standard deviation is $\sqrt{(2 - 18.7)^2 + (3 - 18.7)^2 + \cdots + (36 - 18.7)^2 23} = 8.94$ downloads per hour.

Coefficient of Variation

During the period July 21 to August 7, 2009, the daily closing prices of the Toronto Dominion Bank and the Bank of Nova Scotia had the means and standard deviations given in the following table:

	Mean ($)	Standard Deviation ($)
Toronto Dominion Bank (TD)	62.54	1.62
Bank of Nova Scotia (BNS)	45.36	1.35

The standard deviation for TD is higher than for BNS, but does that mean the share price was more variable? The mean is also higher for TD. If you invested $62.54 in TD, you got a variability in the value of your investment of $1.62. A better measure of variability is the variability per dollar invested. For TD this was 1.62/62.54 = 0.0259. The corresponding figure for BNS was 1.35/45.36 = 0.0297. Per dollar invested, BNS was more variable, even though the standard deviation for TD was higher.

In statistics, we call this the coefficient of variation:

$$\text{CV} = \text{Standard deviation/Mean}$$

$$\text{CV} = s/\bar{y}$$

It measures how much variability exists compared with the mean.

Just Checking

Thinking About Variation

1. Statistics Canada reports the median family income in its summary of census data. Why do you suppose Stats Can statisticians 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 9 litres per 100 kilometres. Of course, yours will "vary." If you had to guess, would you expect the IQR of the fuel efficiency attained by all cars like yours to be: 9, 2, or 0.1 litres per 100 kilometres? Why?
3. A company selling a new MP3 player advertises that the player has a mean lifetime of five years. If you were in charge of quality control at the factory, would you prefer that the standard deviation in lifespans of the players you produce be two years or two months? Why?

L.O. ❷❸

5.5 Reporting the Shape, Centre, and Spread

What should you report about a quantitative variable? Report the shape of its distribution, and include a centre and a spread. But which measure of centre and which measure of spread? The guidelines are straightforward, as described below:

- 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. For unimodal symmetric data, the IQR is usually between one and two standard deviations. 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're 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 centre without a spread can lead you to think you know more about the distribution than you should. Reporting only the spread omits important information.

For Example Summarizing data

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

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 are so unusual as to be considered outliers.

L.O. ❷❸

5.6 Adding Measures of Centre and Spread

We've seen how means and medians are good measures of the centre of a distribution and how IQR, standard deviation, and variance are good measures of spread. This is fine when we have only a single distribution, but often we need more than one. Industrial processes, after all, usually consist of multiple stages. For example, CTS Composites Inc. of Mississauga, Ontario, is a Canadian company that produces metal matrix composites, which are important materials in the automobile industry for disk brakes and are also used in some high-end bicycle frames. Recently it has been found advantageous to manufacture the metal composites in a two-stage production process instead of a single-stage one (www.azom.com/details.asp?ArticleID51383).

Let's suppose we have a two-stage industrial process in which we monitor the processing time taken for 100 products in each stage. The results are given in the table below. We know the centre and the spread for each stage and would like to calculate the corresponding measures for the total time taken.

It's no surprise that we can add the means, but note that we can't add the medians. The mean time in each stage is higher than the median, implying that the distribution is skewed. We know that the median is a natural measure to choose for the centre of a skewed distribution, but we'd need to know how skewed the distributions are in order to calculate the median of the total production time. It can't therefore be done using just the information given. The same is true of the modes: The mode of the total production time can't be calculated as the sum of the modes for each stage. To calculate the median or mode of the total production time, we'd need to go back to the raw data on each of the 100 products.

Processing Time	Number of Products	Mean (minutes)	Median (minutes)	Mode (minutes)	Interquartile Range, IQR (minutes)	Standard Deviation (minutes)	Variance (minutes^2)
Stage 1	100	20	18	17	5	3	9
Stage 2	100	30	26	25	6	4	16
Total	100	50	?	?	?	5 if stages are uncorrelated	25 if stages are uncorrelated

When it comes to calculating measures of spread, we have to be even more careful. The only measure of spread that can be added is the variance, and that can be done only if the times for the two stages are uncorrelated (we'll discuss correlation in a later chapter). Once we've added the variances, we can take the square root of the answer to get the standard deviation of the total production time. The interquartile range for the total production time is like the median and mode: We can't calculate it from the summary statistics for the two stages—we need to know the whole distribution.

L.O. ❷❸

5.7 Grouped Data

Leger Marketing of Montreal asked 1505 Canadians how much extra they would be prepared to pay to purchase a product made in Canada. The results are given in Table 5.2.

Percentage Extra Person Would Be Prepared to Pay	Percentage of Sample
0%	23%
1–5%	14%
6–10%	23%
11–19%	8%
20% or more	17%
No answer	15%

Table 5.2 How much extra Canadians would be prepared to pay to purchase products made in Canada.

Source: Based on Leger Marketing, Montreal. (2003). Canadians and their perceptions on products made in Canada. Retrieved from www.legermarketing.com/documents/SPCLM/031117ENG.pdf

We can't tell from the table the exact extra amount people are prepared to pay; instead, we're given a range—for example, 6–10%. In order to calculate the average percentage that Canadians as a whole are prepared to pay, we base our calculation on the midpoint of the range. The last range given in the table is 20% or more, so we're going to have to *assume* a midpoint for that range—say, 30%. We calculate the mean by multiplying the midpoints by the percentage of people who chose that option and adding the results, as shown in Table 5.3. Our result is that, on average, people are prepared to pay about 8.5% extra to buy Canadian products. This result is only approximate, because some people did not answer the survey and because of our assumption about the 30% midpoint. It's always more accurate to use ungrouped data if available.

Range	Midpoint	% of Sample	MidPt × %
0%	0%	23%	0.00%
1–5%	3%	14%	0.42%
6–10%	8%	23%	1.84%
11–19%	15%	8%	1.20%
>20%	30%	17%	5.10%
		Mean	8.56%

Table 5.3 Calculation of the average extra amount Canadians are prepared to pay in order to buy Canadian products.

The same principle applies to calculating the variance and standard deviation. We use the midpoints of the ranges in our regular formula for variance and also multiply by the percentage, p, of our sample in that group:

$$s^2 = \quad (y - \bar{y})^2 p$$

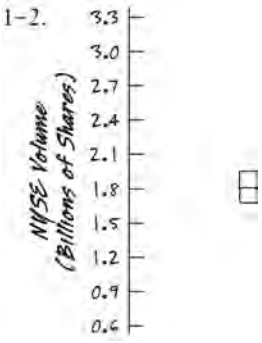
There's no need to divide by n or $n - 1$, since we're working with percentages of the sample, not actual numbers. Once we have the variance, we take its square root to get the standard deviation, as shown in Table 5.4. Note that the standard deviation (SD) is pretty high, due partly to the high percentages of the sample in the lowest and highest categories (23% would pay 0% extra and 17% would pay >20% extra). The coefficient of variation is very high: 10.13/8.56 = 1.18.

Range	Midpoint	% of Sample	MidPt × %	(MidPt – Mean)² × %
0%	0%	23%	0.00%	0.001685
1–5%	3%	14%	0.42%	0.000433
6–10%	8%	23%	1.84%	0.000007
11–19%	15%	8%	1.20%	0.000332
>20%	30%	17%	5.10%	0.007814
		Mean	8.56%	
			Variance =	0.010271
			SD =	10.13%

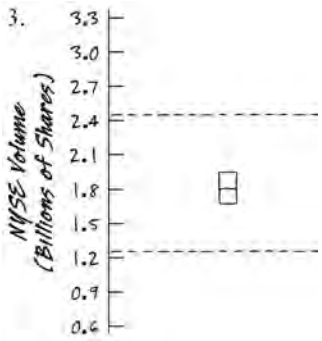
Table 5.4 Calculation of variance and standard deviation for grouped data.

L.O. ❹

5.8 Five-Number Summary and Boxplots

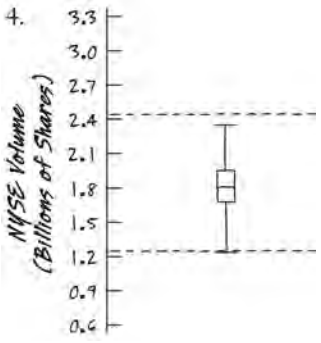


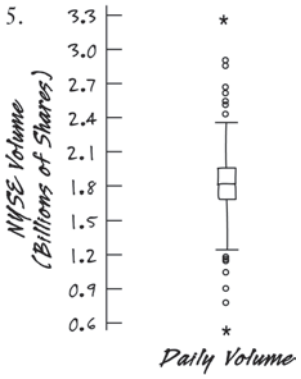
The volume of shares traded on the New York Stock Exchange (NYSE) is important to investors, research analysts, and policy-makers. The volume of shares can predict market volatility, and has been used in models for predicting price fluctuations. How many shares are typically traded in a day on the NYSE? 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). The median and quartiles can be calculated by the methods described in the boxes "Finding the Median by Hand" and "Quartiles" on pages xx and xx. For example, the five-number summary of NYSE volume during the entire year 2006 looks like the values that appear in Table 5.5 (in billions of shares).



Max	3.287
Upper Quartile, Q3	1.972
Median	1.824
Lower Quartile, Q1	1.675
Min	0.616

Table 5.5 The five-number summary of NYSE daily volume (in billions of shares) for the year 2006.





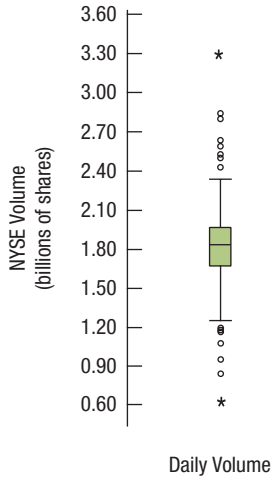


Figure 5.9 Boxplot of daily volume of shares traded on NYSE in 2006 (in billions of shares).

The prominent statistician John W. Tukey, originator of the boxplot, was asked (by one of the authors) why the outlier nomination rule cut at 1.5 IQRs beyond each quartile. His response was that one IQR would be too small and two IQRs would be too large.

The five-number summary provides a good overall summary of the distribution of data. 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 1.675 and 1.972 billion shares. This is the Inter Quartile Range, $IQR = Q3 - Q1 = 0.297$ we can also see the extremes of over 3 billion shares on the high end and just over half a billion 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**. To make a boxplot of the daily volumes, follow these steps:

1. Draw a single vertical axis spanning the extent of the data.
2. Draw short horizontal lines at the lower and upper quartiles and at the median. Then connect them with vertical lines to form a box. The width isn't important unless you plan to show more than one group.
3. Now erect (but don't show in the final plot) "fences" around the main part of the data, placing the upper fence 1.5 IQRs above the upper quartile and the lower fence 1.5 IQRs below the lower quartile. For the NYSE share volume data, compute:

$$\textit{Upper fence} = Q3 + 1.5\ \text{IQR} = 1.972 + 1.5 \times 0.297 = 2.418 \text{ billion shares}$$

and

$$\textit{Lower fence} = Q1 - 1.5\ \text{IQR} = 1.675 - 1.5 \times 0.297 = 1.230 \text{ billion shares}$$

4. Grow "whiskers." Draw lines from each end of the box up and down to *the most extreme data values found within the fences.* If a data value falls outside one of the fences, do *not* connect it with a whisker.
5. Finally, add any outliers by displaying data values that lie beyond the fences with special symbols. In the plot that appears in the margin, about 15 such values exist. (We often use one symbol for outliers that lie less than three IQRs from the quartiles and a different symbol for "far outliers"—data values more than three IQRs from the quartiles.)

Now that you've drawn the boxplot, let's summarize what it shows. The centre of a boxplot is (remarkably enough) a box that shows the middle half of the data, between the quartiles. The height of the box is equal to the IQR. If the median is roughly centred between the quartiles, then the middle half of the data is roughly symmetric. If it's not centred, the distribution is skewed. The whiskers show skewness as well, if they are not roughly the same length. Any outliers are displayed individually, both to keep them out of the way for judging skewness and to encourage you to give them special attention. They may signal mistakes, or they may represent the most interesting cases in your data.

The boxplot for NYSE volume (see Figure 5.9) shows the middle half of the days—those with average volume between 1.676 and 1.970 billion shares—as the central box. From the shape of the box, it looks like the central part of the distribution of volume is roughly symmetric, and the similar length of the two whiskers shows the outer parts of the distribution to be roughly symmetric as well. We also see several high-volume and low-volume days. Boxplots are particularly good at exhibiting outliers. We see two extreme outliers, one on each side. These extreme days may deserve more attention. (When and why did they occur?)

For Example — The boxplot rule for identifying outliers

Question: From the histogram in the first example on page xx, we saw that no points seemed to be so far from the centre as to be considered outliers. Use the 1.5 IQR rule to see if it identifies any points as outliers.

Answer: The quartiles are 13 and 24.5 and the IQR is 11.5. $1.5 \times IQR = 17.25$. 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 Company Customers

In order to focus on the needs of particular customers, companies often segment their customers into groups that display similar needs or spending patterns. A major credit card company wanted to see how much money 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 one-month period last year. Boxplots are especially useful for displaying one variable when combined with a histogram and numerical summaries. Let's summarize the spending of this 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 last year. The data are quantitative, so we'll use histograms and boxplots, as well as numerical summaries.

DO

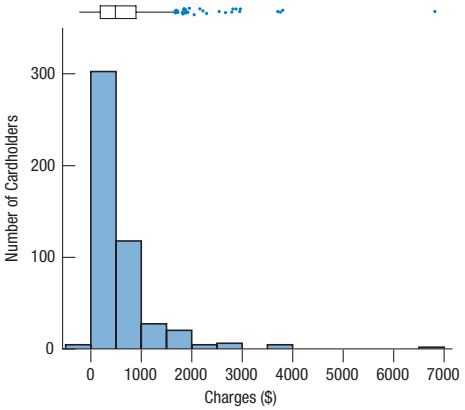
Mechanics Select an appropriate display based on the nature of the data and what you want to know about it.

It's 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 to be in the right ballpark.

The five-number summary of this data is:

Max	6745.01
Q3	738.66
Median	370.65
Q1	114.54
Min	−327.12

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

Count	500
Mean	544.75
Median	370.65
StdDev	661.24
IQR	624.12

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

REPORT

Interpretation Describe the shape, centre, 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 recommend summarizing 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 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.

L.O. ❸

5.9 Percentiles

The box in the middle of the boxplot shows the region between the first quartile, Q1, and the third quartile, Q3, where the centre 50% of the data lies. Twenty-five percent of the data lies below Q1, and another name for Q1 is "25th percentile." Q3 is the 75th percentile. We might also be interested in other percentiles. You can think of **percentiles** as a way of showing where a given percentage of the data lies. For instance, if your mark on this course is at the 82nd percentile, it means that 18% of your classmates got at least as high a mark as you. Notice that 82% is a totally different concept from the 82nd percentile: 82% may be your mark showing what percentage of questions you got right, whereas the 82nd percentile shows how your mark compares with other students' marks.

L.O. ❷❸

5.10 Comparing Groups

As we saw earlier, the volume on the NYSE can vary greatly from 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 might be interested not only in individual daily values, but also in looking for patterns in the volume when we group the days into time periods such as weeks, months, or seasons. Such comparisons of distributions can reveal patterns, differences, and trends.

Let's start with the big picture. We'll split the year into halves: January through June and July through December. Figure 5.10 shows histograms of the NYSE volume for 2006.

The centres and spreads aren't too different, but the shape appears to be slightly right-skewed in the first half, while the second half of the year appears to be left-skewed with more days on the lower end. There are several noticeable outlying values on the high side in both graphs.

Histograms work well for comparing two groups, but what if we want to compare the volume across four quarters? Or 12 months? 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 boxplots 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

Calculating Percentiles

Let us take a simple example of just 12 data values to illustrate the calculation of percentiles. Larger data sets give more accurate results, but they are tough to work with for illustrative purposes. Suppose the number of passengers on 12 flights from Ottawa to Iqaluit is

24, 18, 31, 27, 15, 16,

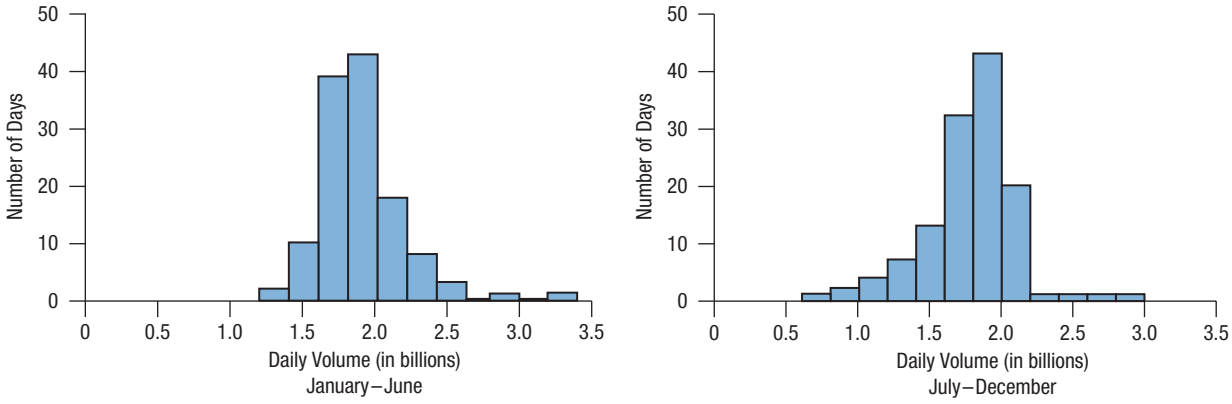
26, 15, 24, 26, 25, 30.

Step 1. We first put the data in ascending order, getting

15, 15, 16, 18, 24, 24,

25, 26, 26, 27, 30, 31.

Step 2: Option 1. Suppose we want to calculate the 80th percentile of this data. Since there are 12 data values, we first calculate 80% of 12, which is 9.6. Since 9.6 *is not* an integer, we round it up to 10 and the 80th percentile is the 10th data value, or 27.



| Figure 5.10 Daily volume on the NYSE split into two halves of the year. How do the two distributions differ?

Step 2: Option 2. Suppose we want to calculate the 50th percentile of the data. We calculate 50% of 12, giving 6. Since 6 *is* an integer, we don't need to round it up. Instead, we take the average of the 6th and 7th data values: (24 + 25)/2 = 24.5. Notice that this follows the same calculation we gave earlier for calculating the median. The median *is* the 50th percentile.

symmetry from whether the medians are centred 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 when 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 — New York Stock Exchange Trading Volume

Are some months on the NYSE busier than others? Boxplots of the number of shares traded by month are a good way to see such patterns. We're interested not only in the centres, but also in the spreads. Are volumes equally variable from month to month, or are they more spread out in some months?

PLAN

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

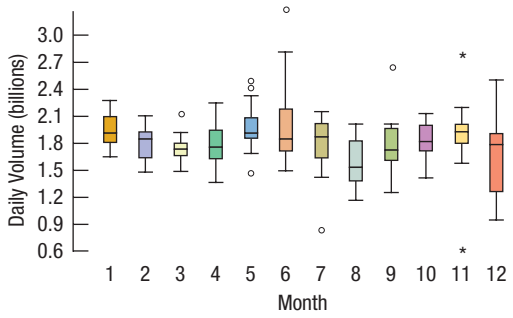
We want to compare the daily volume of shares traded from month to month on the NYSE during 2006.

The daily volume is quantitative and measured in number of shares.

DO

Mechanics Choose an appropriate display for the data.

We can partition the values by month and use side-by-side boxplots to compare the volume across months.



REPORT LO 1

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

MEMO

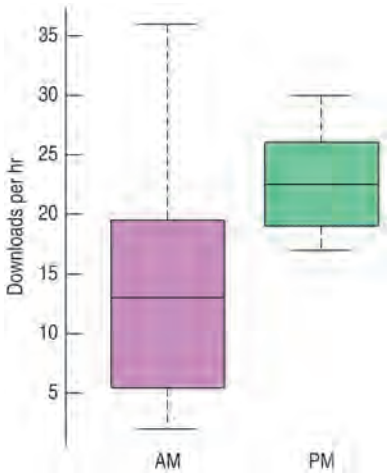
Re: Research on the Trading Volume of the NYSE

We have examined the daily sales volume on the NYSE (number of shares traded) for each month of 2006. As the attached display shows, sales volume has lower median volume in March and August. The highest median trading activity is found in November. The variability of trading volume also shows a pattern. June and December have higher variability than the rest, and March has noticeably less variability. There were several unusually high-volume days that bear investigation and extremely low-volume days in July and November.

For Example Comparing boxplots

Question: For the data in the first example on page xx, compare the A.M. downloads with the P.M. downloads by displayings 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 P.M. downloads are also much more consistent. The entire range of the P.M. hours, 15, is about the size of the IQR for A.M. hours. Both distributions appear to be fairly symmetric, although the A.M. hour distribution has some high points, which seem to give some asymmetry.



L.O. ❺

5.11 Dealing with Outliers

When we looked at a boxplot for NYSE trading volumes of the entire year, there were 15 outliers. Now, when we group the days by *Month*, the boxplots display fewer days as outliers, and identify different days as the extraordinary ones. This change occurs because our outlier nomination rule for boxplots depends on the quartiles of the data being displayed. Days that may have seemed ordinary when placed against the entire year's data can look like outliers for the month they're in, and vice versa. That high-volume day in March certainly wouldn't stand out in May or June, but for March it was remarkable, and that very low-volume day in November really stands out now. What should we do with such outliers?

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 (those more than 1.5 IQRs outside the box), but that's just a rule of thumb—not a definition. Also, the rule doesn't tell you what to do with them.

So, what *should* we do with outliers? The first thing to do is to try to understand them in the context of the data. Look back at the boxplot in the Guided Example. The boxplot for November (month 11) shows a fairly symmetric body of data, with one low-volume day and one high-volume day set clearly apart from the other days. Such a large gap suggests that the volume really is quite different.

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. Or the units may be wrong. If you saw the number of shares traded on the NYSE listed as two shares for a particular day, you'd know something was wrong. It could be that it was meant as 2 billion shares, but you'd have to check to be sure. Sometimes a number is transcribed incorrectly, perhaps copying an adjacent value on the original data sheet. If you can identify the error, then you should certainly correct it.

Many outliers are not incorrect; 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 data set.

What about that low November day? It was November 24, 2006, the Friday after the American Thanksgiving, a day when, most likely, traders would have rather stayed home.

The high-volume day, September 15, was a "triple witching day"—a day when, during the final trading hour, options and futures contracts expire. Such days often experience large trading volume and price fluctuations.

Increase in 14-Year-Old Widowers?

Careful attention to outliers can often reveal problems in data collection and management. Two researchers, Ansley Coale and Fred Stephan, looking at data from the 1950 U.S. census, noticed that the number of widowed 14-year-old boys had increased from 85 in 1940 to a whopping 1600 in 1950. The number of divorced 14-year-old boys had increased, too, from 85 to 1240. Oddly, the number of teenaged widowers and divorcees *decreased* for every age group after 14, from 15 to 19. When Coale and Stephan also noticed a large increase in the number of young Native Americans in the Northeast United States, they began to look for data problems. As it turns out, data in the 1950 census were recorded on computer cards. Cards are hard to read and mistakes are easy to make. It turned out that data punches had been shifted to the right by one column on hundreds of cards. Because each card column meant something different, the shift turned 43-year-old widowed males into 14-year-olds, 42-year-old divorcées into 14-year-olds, and children of white parents into Native Americans. Not all outliers have such a colourful (or famous) story associated with them, but it's always worthwhile to investigate them. And, as in this case, the explanation is often surprising. (**Source:** Based on Coale, A., & Stephan, F. [1962, June]. The case of the Indians and the teen-age widows. *Journal of the American Statistical Association*, 57, 338–347.)

For example Dealing with outliers and summarizing data

Question: A real estate report lists the following prices for sales of single-family homes in a small town in Alberta (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

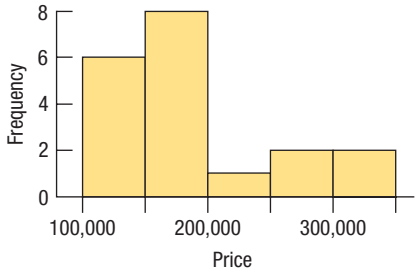
Answer: A boxplot shows an extreme outlier:

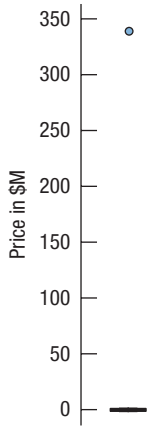
That extreme point is a home whose sale price is listed at $339.4 million.

A check on the Internet shows that the most expensive homes ever sold are less than $300 million.

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 and $212,500 with an IQR of $68,500.

L.O. ❻

5.12 Standardizing

The data we compared by groups in previous sections all represented the same variable. It was easy to compare volume on the NYSE in July with volume on the NYSE in December because the data had the same units. Sometimes, however, we want to compare very different variables—apples to oranges, so to speak. For example, the Great Place to Work Institute in the U.S. measures more than 50 aspects of companies and publishes, through *Fortune* magazine, a ranking of the top places to work in that country. In 2007, the top honour was won by Google.

What was the key to Google's winning? Was it the free food offered to all employees? Maybe the on-site day care? How about the salaries—do they compare favourably with those of other companies? Did they score better on all 50 variables? Probably not, but it isn't obvious how to combine and balance all these different aspects to come up with a single number. The variables don't even have the same units; for example, average salary is in dollars, perceptions are often measured on a seven-point scale, and diversity measures are in percentages.

The trick to comparing very different-looking values is to standardize them. Rather than working with the original values, we ask, "How far is this value from the mean?" Then—and this is the key—we measure that distance with the standard deviation. The result is the standardized value, which records how many standard deviations each value is above or below the overall mean. The standard deviation provides a ruler, based on the underlying variability of all the values, against which we can compare values that otherwise have little in common.

It turns out that statisticians do this all the time. Over and over during this course (and in any additional Statistics courses you may take), questions such as "How far is this value from the mean?" and "How different are these two values?" will be answered by measuring the distance or difference in standard deviations.

In order to see how standardizing works, we'll focus on just two of the 50 variables that the Great Places to Work Institute reports—the number of *New Jobs* created during the year and the reported *Average Pay* for salaried employees—for two companies. We'll choose two companies that appeared in ranking farther down the list to show how standardization works: Starbucks and the Wrigley Company (the company that makes Wrigley's chewing gum, among other things).[5]

When we compare two variables, it's always a good idea to start with a picture. Here we'll use stem-and-leaf displays (Figure 5.11) so that we can see the individual distances, highlighting Starbucks in red and Wrigley in blue. The mean number of new jobs created for all the companies was 305.9. Starbucks, with over 2000 jobs, is well above average, as we can see from the stem-and-leaf display. Wrigley, with only 16 jobs (rounded to 0 in the stem-and-leaf), is closer to the centre. On the other hand, Wrigley's average salary was $56,350 (rounded to 6), compared with Starbucks' $44,790 (represented as 4), so even though both are below average, Wrigley is closer to the centre (see brief table in margin).

Variable	Mean	SD
New Jobs	305.9	1507.97
Aug. Pay	$73,299.42	$34,055.25

When we compare scores from different variables, our eye naturally looks at how far from the centre of each distribution the value lies. We adjust naturally for the fact that these variables have very different scales. Starbucks did better on *New Jobs*, and Wrigley did better on *Average Pay*. To quantify *how much* better each one

[5]The data we analyze here are actually from 2005, the last year for which we have data, and the year Wegman's Supermarkets was the number one company to work for.

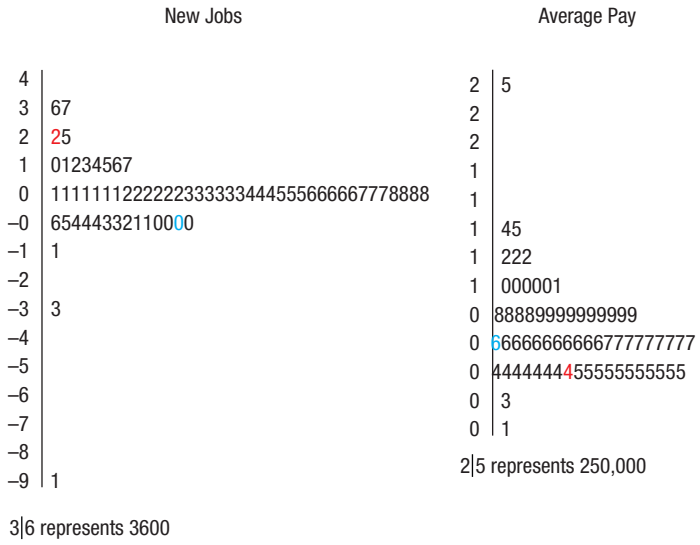


Figure 5.11 Stem-and-leaf displays for both the number of *New Jobs* created and the *Average Pay* of salaried employees at the top 100 companies to work for in 2005 from *Fortune* magazine. Starbucks (in red) created more jobs, but Wrigley (in blue) did better in average pay. Which company did better for both variables combined?

did and to combine the two scores, we'll ask how many standard deviations they each are from the means.

To find how many standard deviations a value is from the mean, we find

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

We call the resulting value a **standardized value** and denote it z. Usually, we just call it a z**-score**.

A z-score of 2.0 indicates that a data value is two standard deviations above the mean. Data values below the mean have negative z-scores, so a z-score of 0.84 means that the data value is 0.84 standard deviations *below* the mean. A rule of thumb for identifying outliers is $z > 3$ or $z < -3$.

	New Jobs	Average Pay
Mean	305.9	$73,299.42
(all companies) SD	1507.97	$34,055.25
Starbucks	2193	$44,790
***z*-score**	**1.25** = (2193 − 305.9)/1507.9716	**0.84** = (44790 − 73299.42)/34055.25 $56,351
Wrigley *z*-score	**0.19** = (16 − 305.9)/1507.97	**0.50** = (56351 − 73299.42)/34055.25

Table 5.6 For each variable, the z-score for each observation is found by subtracting the mean from the value and then dividing that difference by the standard deviation.

Standardizing into *z*-Scores

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

Starbucks offered more new jobs than Wrigley, but Wrigley had a higher average salary (see Table 5.6). 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.

For Example Comparing values by standardizing

Question: A real estate analyst finds more data from home sales, as discussed in the previous example on page xxx. 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 500 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.

L.O. ❼

5.13 Time Series Plots

The volume on the NYSE is reported daily. Earlier, we grouped the days into months and half-years, but we could simply look at the volume day by day. Whenever we have time series data, it is a good idea to look for patterns by plotting the data in time (sequential) order. Figure 5.12 shows the *Daily Volumes* plotted over time for 2006.

A display of values against time is sometimes called a **time series plot**. This plot reflects the pattern that we saw when we plotted the daily volume by month, but without the arbitrary divisions between months, we can see periods of relative calm contrasted with periods of greater activity. We can also see that the volume both became more variable and increased during certain parts of the year.

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 (see Figure 5.13).

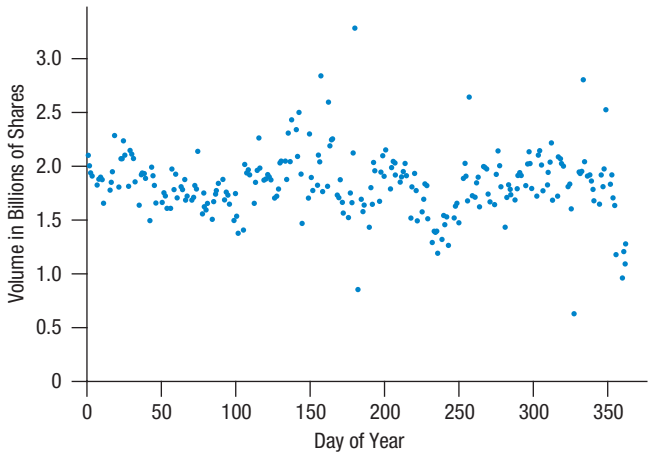


Figure 5.12 A time series plot of *Daily Volume* shows the overall pattern and changes in variation.

It's often 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 to think about how the values vary around that trend—the time series version of centre 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

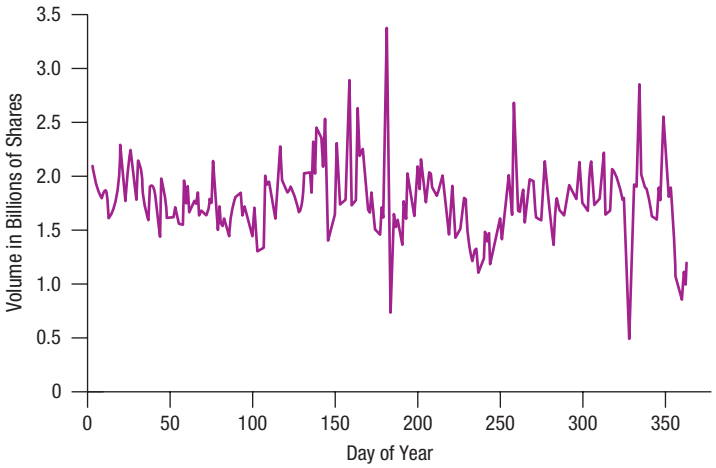


Figure 5.13 The *Daily Volumes* of Figure 5.12, drawn by connecting all the points. Sometimes this can help us see the underlying pattern.

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 represents the daily volumes of Figures 5.12 and 5.13 with a typical smoothing function, available in many statistics programs.[6] 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.

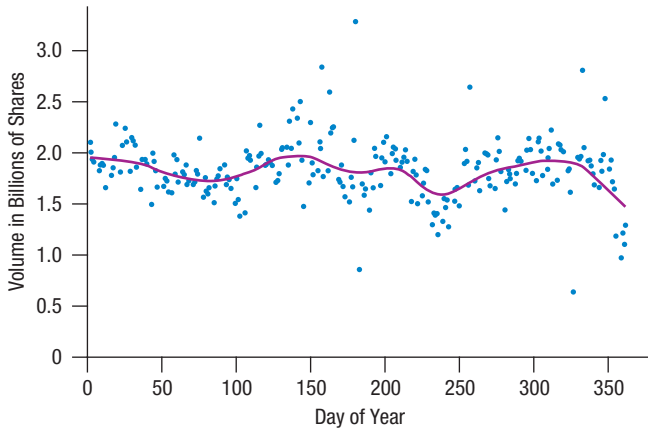


Figure 5.14 The *Daily Volumes* of Figure 5.12, with a smooth trace added to help your eye see the long-term pattern.

It's 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 and less activity during the week between Christmas and New Year's Day. But we certainly wouldn't predict a record every June 30.

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, and at a given speed and direction, is well understood. The path interest rates will take is much less clear.

[6]We discuss the main ways to smooth data in Chapter 22 .

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 to this phenomenon later in this book to better understand when, how, and how much we can justify doing that.

Let's return to the Bell Canada data we saw at the beginning of the chapter. The stock price changes are a time series from September 12 to October 21, 2011. The histogram (Figure 5.1) showed a roughly symmetric, unimodal distribution for the most part concentrated between −$0.20 and +$0.10, but it doesn't show whether the pattern changes over time. The time series plot in Figure 5.15 shows a different story.

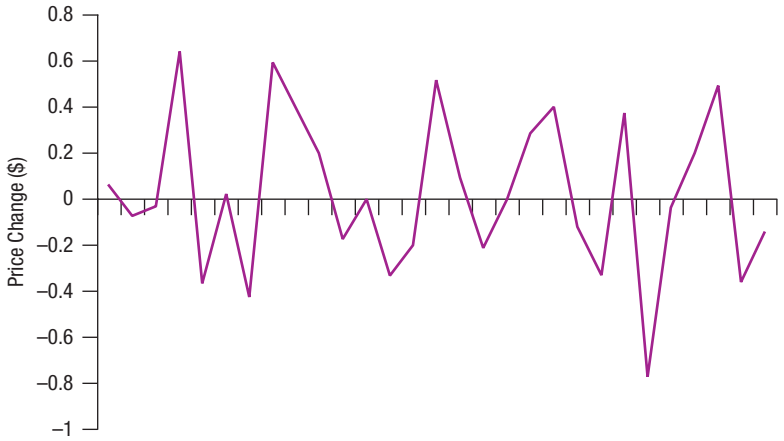
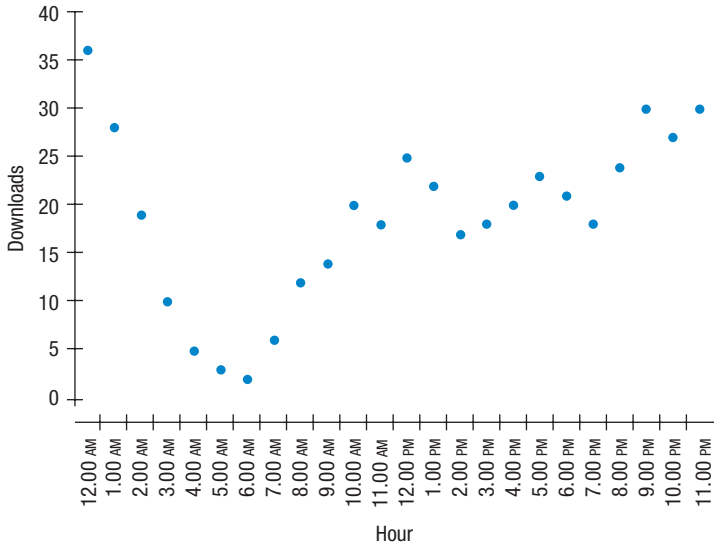


Figure 5.15 A time series plot of daily Bell Canada stock price changes.

For Example Plotting time series data

Question: The downloads from the first example on page xx 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 A.M. 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. If we'd represented this data using a histogram, we would have missed this pattern entirely.



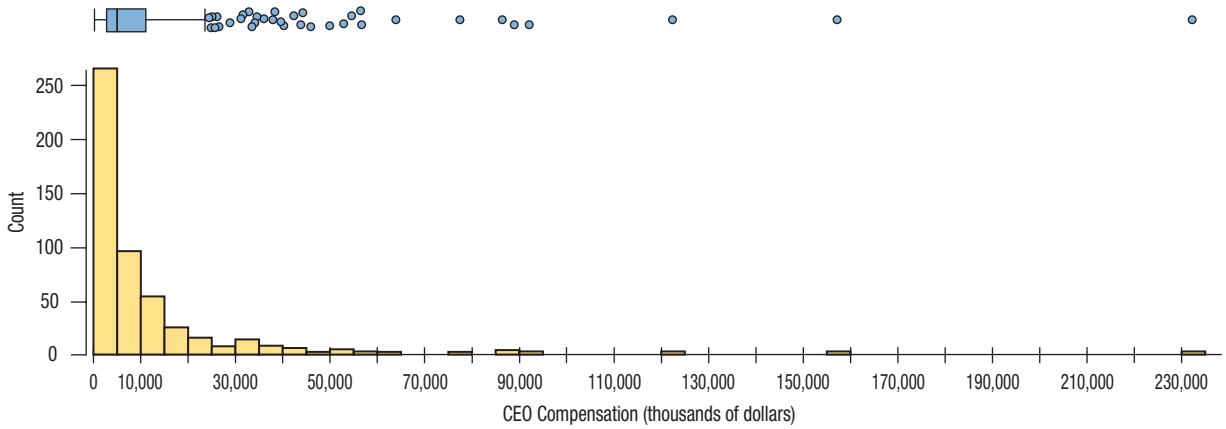
The time series plot of the Bell Canada stock price changes shows the same variability as was shown by the histogram; it also shows that this pattern is pretty constant throughout the length of the data series. A slight downward trend in the average level of the data is apparent as well. A time series that does *not* change over time is called **stationary**. Our data have a stationary variability and a slightly non-stationary average level. When a data series is very non-stationary, a time series plot is a better graphical representation than a histogram.

L.O. ❷

*5.14 Transforming Skewed Data

When a distribution is skewed, it can be hard to summarize the data simply with a centre 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.[7] What does the distribution of the Fortune 500 companies' CEOs look like? Figure 5.16 shows 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 why 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 the boxplot make it clear that this distribution is *very* skewed to the right.

- **Dealing with logarithms.** You probably don't encounter logarithms every day. In this text, 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 three 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

[7] **Sources: Based on** 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.

$\log_{10}$ 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 computer button.

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 "centre" 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.

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.

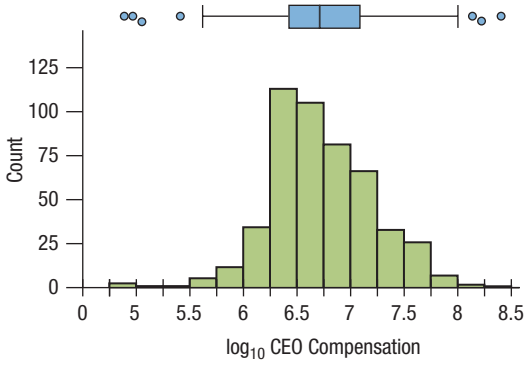


Figure 5.17 Taking logs makes the histogram of CEO total compensation nearly symmetric.

Math Box

Optional

Let's calculate the mean of the data after taking logs.

If C_n is the compensation of the n^{th} CEO, then:

$\text{Log (Mean)} = [\ \text{Log}(C_1) + \text{Log}(C_2) + \cdots + \text{Log}(C_n)\ ]/n$

$= \text{Log}\ [C_1 \times C_2 \times \ldots \times C_n\ ] / n$

$= \text{Log}\ [\ (C_1 \times C_2 \times \ldots \times C_n)^{1/n}\ ]$

$\text{Mean} = (C_1 \times C_2 \times \ldots \times C_n)^{1/n}$

This type of mean is what we called the "geometric mean" in Section 5.3.

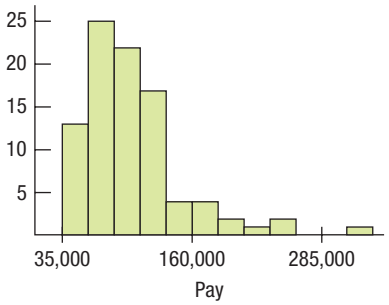
When we re-express the compensation of CEOs by taking logs, we end up with a histogram in which the data are more grouped together, which is useful from

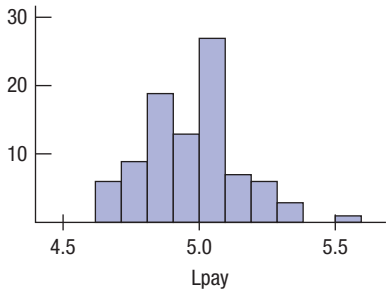
the standpoint of getting a clear *graphical* representation of the data. Figure 5.17 is easier on the eyes than Figure 5.16. This does *not* imply that the mean of Figure 5.17 is somehow a "better" way of measuring the centre of the data than the mean of Figure 5.16. Each mean is valid so long as we bear in mind what it is the mean of—either the CEO compensation or the log of the CEO compensation. In fact, the Math Box shows that the CEO compensation from calculating the mean of Figure 5.17 is the same as the geometric mean of the original data. It's just another way of calculating the mean. Neither way is right or wrong.

A major advantage of re-expressing or transforming data comes when we make inferences about our data using the statistical tests described in Part 3 of this book. Most of those tests work better when the data have a symmetric, bell-shaped distribution. No data are ever going to be perfectly symmetric or bell-shaped, but the transformed CEO compensation in Figure 5.17 is certainly more amenable to these methods of statistical inference than the raw data in Figure 5.16. Chapter 17," The Nonparametric Methods," provides methods that can be used on non-symmetric data.

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's 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. Figure 5.18 is a histogram of the insurance policy numbers of some workers. It's not very informative because the policy numbers are categorical. Generating a histogram or stem-and-leaf display of a categorical variable makes no sense. A bar chart or pie chart may do better.

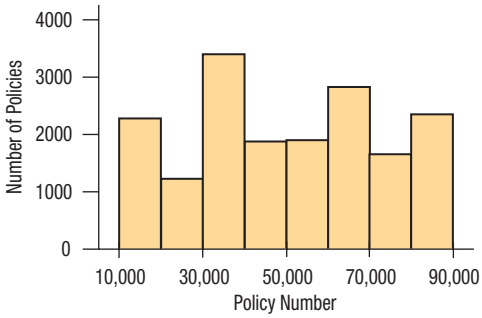


Figure 5.18 It's not appropriate to display categorical data like policy numbers with a histogram.

- **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. If you're not using software with these features, you can always use around $\log_2 n$ bins. Bear in mind, though, that using too many bins can result in a random-looking histogram, and using too few bins can result in a loss of detail.
- **Avoid inconsistent scales.** Parts of displays should be mutually consistent. It's not fair to change scales in the middle or plot two variables on different scales within 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 labelled so that readers can understand what the plot displays.

Here's a remarkable example of a plot gone wrong. It illustrated a news story about rising college costs in the U.S. 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.

- **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 beginning in 1989. Plotting the latter on the same (invisible) scale as the former makes it seem that they're for the same years.
- **The vertical axis isn't labelled.** 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 data display 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's in the centre of the histogram? Think about the spread. An IQR of 20 litres per 100 kilometres 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 employee number or the standard deviation of social insurance numbers is not meaningful. If the variable is categorical, you should instead report summaries such as percentages. It's 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 can't separate the data in a meaningful way, you shouldn't summarize the centre 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 can't do either of these, then choose the median and IQR to summarize the centre and spread.

Ethics in Action

Beth Ghazi owns Zenna's Café, an independent coffee shop located in a small city in Atlantic Canada. 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 of the best in the area. In addition to providing the highest-quality coffees, Beth wants her business to be socially responsible. To that end, she pays fair prices to coffee farmers and donates profits to help charitable causes in Panama, Costa Rica, and Guatemala. She also 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 message of social responsibility to the public, but with an emphasis on its 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 vol-

unteered 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 that 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 Ghazi and Zenna's Café (related to Item C, ASA Ethical Guidelines; see Appendix C, the American Statistical Association's Ethical Guidelines for Statistical Practice, also available online at www.amstat.org/about/ethicalguidelines.cfm).*

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 (and 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

❶ We've learned how to display and summarize quantitative data to help us see the story the data have to tell.

- We can display the distribution of quantitative data with a histogram or a stem-and-leaf display.
- We've seen the power of transforming our data so that it's not so skewed.

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

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

❸ Measures of spread include the range, IQR, and standard deviation.

- We'll report the median and IQR when the distribution is skewed. If it's symmetric, we'll summarize the distribution with the mean and standard deviation. Always pair the median with the IQR and the mean with the standard deviation.
- We've seen how to calculate percentiles and how to use them, particularly with skewed data.

❹ We've learned the value of comparing groups and looking for patterns among groups and over time.

- We've seen that boxplots are very effective for comparing groups graphically.
- When we compare groups, we discuss their shapes, centres, spreads, and any unusual features.

❺ We've experienced the value of identifying and investigating outliers, and we've seen that when we group data in different ways, it can allow different cases to emerge as possible outliers.

❻ We've learned the power of standardizing data.

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

❼ We've graphed data that have been measured over time against a time axis and looked for trends both by eye and with a data smoother.

Terms

Bimodal distributions Distributions with two modes.

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

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

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

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

Geometric mean A measure of the centre of a set of data $a_1, a_2, \ldots, a_n$, given by: $(a_1 \times a_2 \times \cdots \times a_n)^{1/n}$

Histogram (relative frequency) 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 centre found as $\frac{\sum y}{n}$.

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 data distribution. The apparent location of modes can change as the scale of a histogram is changed.

Multimodal distributions 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 just by looking at the numbers. We need to probe further and find out where the numbers came from.

Percentile A value below which a given percentage of data lies. For instance, 10% of data is below the 10th percentile.

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 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 We re-express or transform data by taking the logarithm, square root, reciprocal, or some other mathematical operation on all values of the data set.

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 centre. Measures of spread include the IQR and the standard deviation.

Standard deviation A measure of spread found as $s = \sqrt{\frac{(y - \bar{y})^2}{n - 1}}$ for sample data, and $\sigma = \sqrt{\frac{(y - \mu)^2}{n}}$ for population data.

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 data distribution is approximately symmetric if the two halves on either side of the centre 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 data distribution that's roughly flat is said to be approximately 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.

Skills

Plan

- Be able to identify an appropriate display for any quantitative variable.
- Be able to select a suitable measure of centre and a suitable measure of spread for a variable based on information about its distribution.
- Know the basic properties of the median: The median divides the data into the half of the data values that are below the median and the half that are above the median.
- Know the basic properties of the mean: The mean is the point at which the histogram balances.
- Know that the standard deviation summarizes how spread out all the data are around the mean.
- Know that standardizing uses the standard deviation as a ruler.
- Know how to display the distribution of a quantitative variable with a stem-and-leaf display or a histogram.

Do

- Know how to make a time series plot of data that are collected at regular time intervals.
- Know how to compute the mean and median of a set of data and know when each is appropriate.
- Know how to compute the standard deviation and IQR of a set of data and know when each is appropriate.
- Know how to compute a five-number summary of a variable.
- Know how to calculate percentiles.
- Know how to construct a boxplot by hand from a five-number summary.
- Know how to calculate the z-score of an observation.

Report

- Be able to describe and compare the distributions of quantitative variables in terms of their shape, centre, and spread.
- Be able to discuss any outliers in the data, noting how they deviate from the overall pattern of the data.
- Be able to describe summary measures in a sentence. In particular, know that the common measures of centre and spread have the same units as the variable they summarize and that they should be described in those units.
- Be able to compare two or more groups by comparing their boxplots.
- Be able to discuss patterns in a time series plot, in terms of both the general trend and any changes in the spread of the distribution over time.

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's 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.

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

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

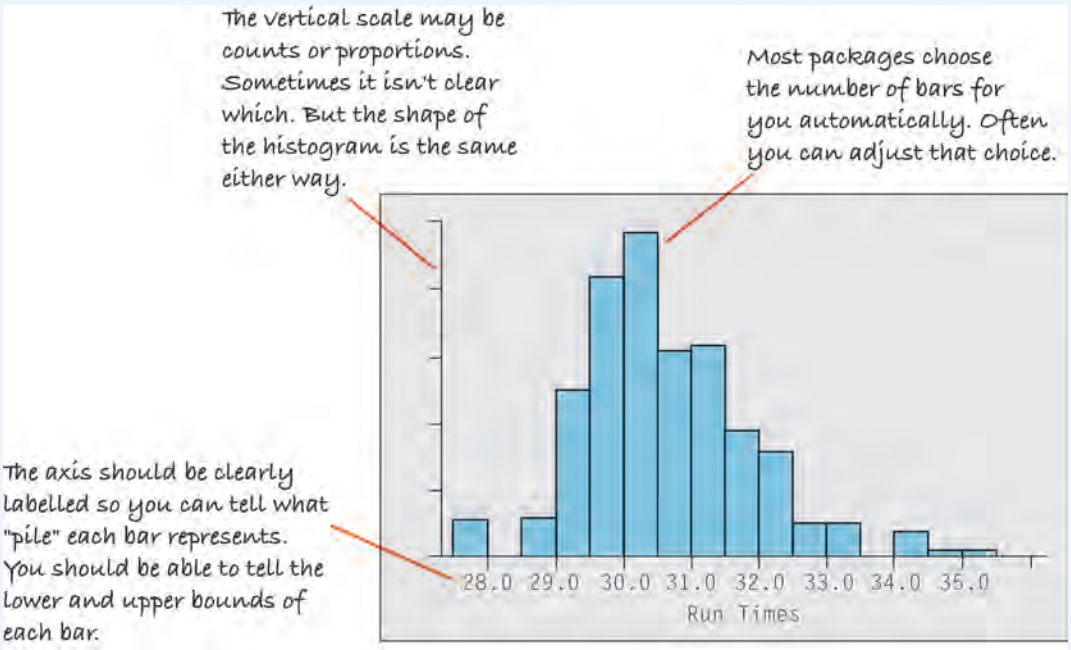
Most packages choose the number of bars for you automatically. Often you can adjust that choice.

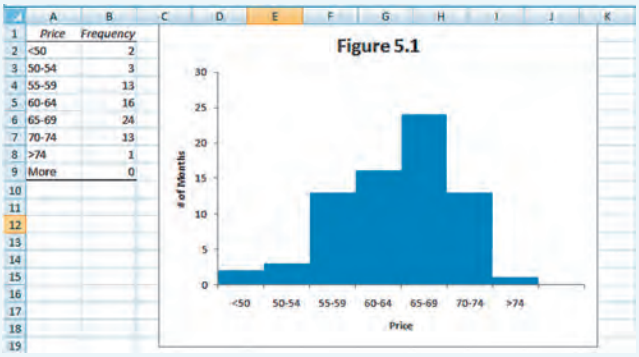
EXCEL

*x*LSTAT

To make a histogram in Excel 2007 or 2010, use the Data Analysis add-in. If you haven't installed it, 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 dialogue 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 tailing within each bin using an Excel function such as FREQUENCY (Data array, Bins array). Consult your Excel manual or help files for details on how to do this.

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 labelled **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** dialogue, assign one or more variables from the source list to the Dependent List and click the **OK** button.

JMP

To make a histogram and find summary statistics,

- Choose **Distribution** from the **Analyze** menu.
- In the **Distribution** dialogue, 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.

MINI case studies

Canadian Exports

Statistics on Canadian exports are used for a variety of purposes, from projecting Canada's foreign exchange earnings to planning capacity in Canadian ports. The file **ch05_MCSP_Canadian_Exports** contains monthly export data from Statistics Canada for four selected products: wheat, zinc, fertilizer, and industrial machinery. Statistics Canada calculates exports on a "Customs" basis and also on a "Balance of Payments" basis, and the file contains footnotes describing the difference.[8]

a) Draw time series graphs of this export data and identify any major differences between the "Customs" and "Balance of Payments" series.

b) Explain which basis of calculation, "Customs" or "Balance of Payments," would be appropriate for projecting Canada's foreign exchange earnings.

c) Explain which basis of calculation, "Customs" or "Balance of Payments," would be appropriate for planning capacity in Canadian ports.

d) Are there any exceptional periods during which exports of wheat, zinc, fertilizer, or industrial machinery have differed from overall trends?

Hotel Occupancy Rates

Many properties in the hospitality industry experience strong seasonal fluctuations in demand. To be successful in this industry, it's important to anticipate such fluctuations and to understand demand patterns. The file **ch05_MCSP_Occupancy_Rates** contains data on quarterly *Hotel Occupancy Rates* (in % capacity) for a town in southern Ontario from January 2000 to December 2007.

Examine the data and prepare a report for the manager of a hotel in the town in southern Ontario 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 in how aggressive they'd 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

[8]**Source:** Based on Statistics Canada. (2012). CANSIM using CHASS, Table 228-0001: Merchandise imports and exports, by major groups and principal trading areas for all countries, monthly (dollars).

consistent performance and are expected to continue to do well. Value stocks, on the other hand, are stocks whose prices seem low compared with 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.[9]

In the data set **ch05_MCP_Returns**[10] 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.

MyStatLab Visit the MyStatLab website at www.pearsoned.ca/mystatlab. This online homework and tutorial system puts you in control of your own learning with study and practice tools directly correlated to this chapter's contents.

[9] The cynical statistician might say that the manager who invests in growth funds puts his faith in extrapolation, while the value manager is putting his faith in the Law of Averages.

[10] 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 customer 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. **L.O. ❶**

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 (in Canadian dollars) from yesterday's records:

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. **L.O. ❶**

SECTION 5.2

3. For the histogram you made in Exercise la,
a) Is the distribution unimodal or multimodal?
b) Where is (are) the mode(s)?
c) Is the distribution symmetric?
d) Are there any outliers? **L.O. ❷**

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? **L.O. ❷**

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. **L.O. ❷**

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. **L.O. ❷**

SECTION 5.4

7. For the data in Exercise 1:

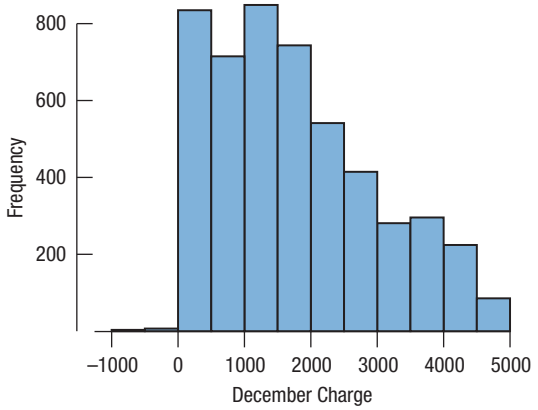
a) Find the quartiles using your calculator.
b) Find the quartiles using the method in the "Quartiles" box on page xx.
c) Find the IQR using the quartiles from part b.
d) Find the standard deviation. **L.O. ❸**

8. For the data in Exercise 2:
a) Find the quartiles using your calculator.
b) Find the quartiles using the method in the "Quartiles" box on page xx.
c) Find the IQR using the quartiles from part b.
d) Find the standard deviation. **L.O. ❸**

SECTION 5.5

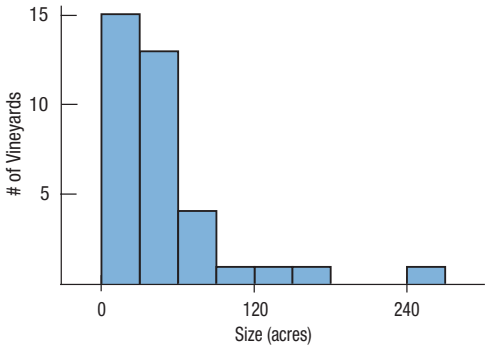
9. The histogram shows the December charges (in $) for 5000 customers in one marketing segment of a credit card company. (Negative values indicate customers who received more credits than charges during the month.)

a) Write a short description of this distribution (shape, centre, 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 centre, the mean, or the median? Explain. **L.O. ❷, ❸**



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 the managers are considering an expansion of their relatively small establishment, they're curious about how their size compares to that 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, centre, 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 centre, the mean, or the median? Explain. **L.O. ❷, ❸**



SECTION 5.6

11. The spending in dollars of 26,790 customers in one marketing segment of a credit card company, during June and July last year, is summarized in the table below.

	Mean	First Quartile	Median	Third Quartile	Standard Deviation
June	876	328	731	1658	986
July	793	387	798	1980	1298

If possible, calculate the mean, median, inter-quartile range, and standard deviation for the total spending of these customers for June plus July. State any assumptions you make. **L.O. ❷, ❸**

12. In order to get to campus, a student has to walk to the bus stop and then take a bus to the university. She monitors how much time this journey takes for 55 days. The time taken in minutes for each stage of her journey varies according to the information in the following table.

	Mean	Median	Interquartile Range	Standard Deviation
Walk	11	10	3	2
Bus	14	12	4	3

If possible, calculate the mean, median, interquartile range, and standard deviation of the total travel time. State any assumptions you make. **L.O. ❷, ❸**

SECTION 5.7

The table below gives the age distribution of the Canadian population according to the 2006 Census.

	Male	Female
Under 5 years	864,600	825,940
5 to 9 years	926,860	882,515
10 to 14 years	1,065,865	1,014,065
15 to 24 years	2,143,235	2,077,645
25 to 34 years	1,963,660	2,042,145
35 to 44 years	2,369,030	2,449,705
45 to 54 years	2,449,095	2,528,805
55 to 64 years	1,806,530	1,867,960
65 to 74 years	1,087,270	1,201,095
75 to 84 years	637,905	888,375
85 years and over	161,920	358,685
Total	15,475,970	16,136,935

Source: Statistics Canada. 2006 Census of population. Statistics Canada catalogue no. 97-551-XCB2006005 (Canada, Code 01).

13. Calculate the average age of males in the Canadian population in 2006, assuming that the average age of people under 5 is 2.5 and that the average age of people over 85 is 92. **L.O. ❷, ❸**

14. Calculate the average age of females in the Canadian population in 2006, assuming that the average age of people under 5 is 2.5 and that the average age of people over 85 is 92. **L.O. ❷, ❸**

SECTION 5.8

15. 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? **L.O. ❹**

16. 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? **L.O. ❹**

17. 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? **L.O. ❹**

18. A survey of major universities asked what percentage of incoming students usually graduate "on time" in four 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. **L.O. ❹**

SECTION 5.9

19. Calculate the 10th and 80th percentile of the ages of the customers in Exercise 1. Interpret the meaning of the 80th percentile. **L.O. ❸**

20. Calculate the 12th and 60th percentile of the purchases of the customers in Exercise 2. Interpret the meaning of the 12th percentile. **L.O. ❸**

SECTION 5.10

21. The survey from Exercise 1 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. **L.O. ❷, ❸**

22. The store manager from Exercise 2 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

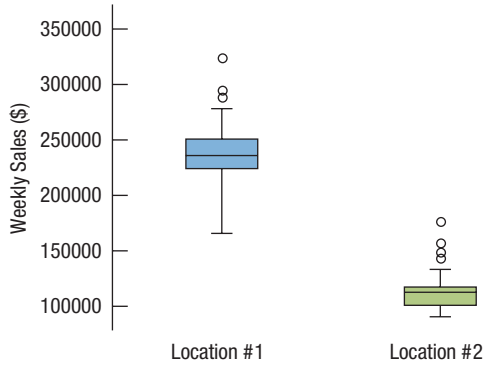
Weekend $n = 150$

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

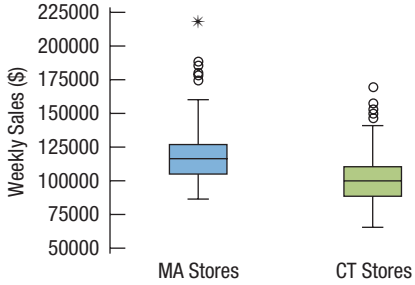
From these statistics, construct side-by-side boxplots and write a sentence comparing the two distributions. **L.O. ❷, ❸**

SECTION 5.11

23. Here are boxplots of the weekly sales 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 and 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. **L.O. ❺**



24. Recall the distributions of the weekly sales for the regional stores in Exercise 23. Following are boxplots of weekly sales for this same food store chain for three stores of similar size and location for two different provinces: Alberta (AB) and Saskatchewan (SK). Compare the distribution of sales for the two provinces and describe in a report. **L.O. ❺**



SECTION 5.12

25. 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? **L.O. ❻**

26. 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 would a purchase with a z-score of 3.5 be? **L.O. ❻**

SECTION 5.13

The table below gives the percentage of the Ontario and B.C. population aged 65 years and older according to the Canadian Census from 1956 to 2006.

	ON	B.C.
1956	8.4	10.8
1961	8.1	10.2
1966	8.2	9.5
1971	8.4	9.4
1976	8.9	9.8
1981	10.1	10.9
1986	10.9	12.1
1991	11.7	12.9
1996	12.4	12.8
2001	12.9	13.6
2006	13.6	14.6

27. Draw a time series plot of the percentage of the Ontario population aged 65 years and older from 1956 to 2006. Describe the trends that emerge. **L.O. ❼**

28. Draw a time series plot of the percentage of the B.C. population aged 65 years and older from 1956 to 2006. Describe the trends that emerge. **L.O. ❼**

SECTION 5.14

29. 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?

30. 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

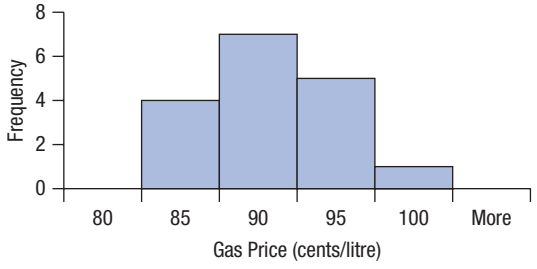
31. Statistics in business. Find a histogram that shows the distribution of a variable as it appeared in a business publication (e.g., *The Wall Street Journal*, *National Post*, *The Economist*, etc.).

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. **L.O. ❶**

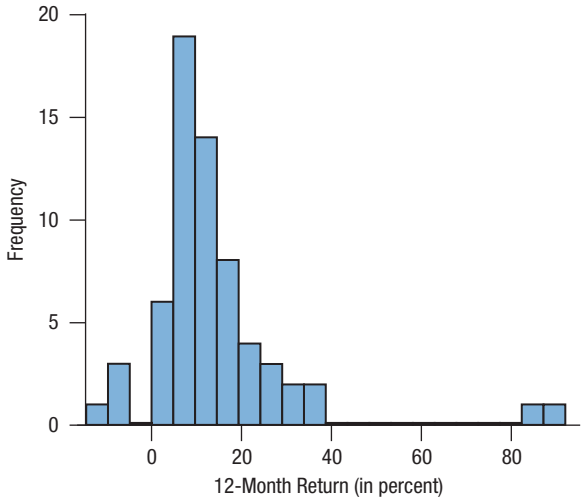
32. Statistics in business, part 2. Find a graph other than a histogram that shows the distribution of a quantitative variable as it appeared in a business publication (e.g., *The Wall Street Journal*, *The Globe and Mail*, *The Economist*, etc.).

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. **L.O. ❼**

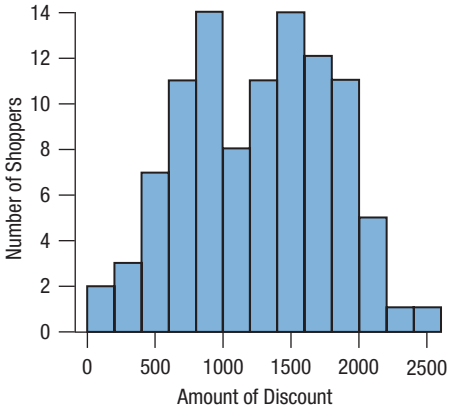
33. Ottawa gas prices. The histogram below shows the price of regular gas at 17 gas stations on a specific day in 2009. For instance, the bar at 85 indicates that there were four gas stations with prices between 80 and 85 cents/litre. Describe the shape of the distribution and its centre and spread. **L.O. ❶, ❷, ❸**



34. Mutual funds. The histogram below displays the 12-month returns (in percent) for a collection of mutual funds in 2007. Give a short summary of this distribution (shape, centre, spread, unusual features). **L.O. ❶, ❷, ❸**



35. 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, centre, spread, unusual features). What do you think might account for this particular shape? **L.O. ❶, ❷, ❸**



36. Mutual funds, part 2. Use the data set in Exercise 34 to answer the following questions.

a) Find the five-number summary for these data.
b) Find appropriate measures of centre 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? **L.O. ❶, ❷, ❸, ❹**

37. Car discounts, part 2. Use the data set in Exercise 35 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 35 that isn't clear in the boxplot? **L.O. ❶, ❷, ❸, ❹**

38. Hockey. During his 20 seasons in the National Hockey League, from 1979 to 1998, 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? **L.O. ❶, ❷, ❸, ❹**

39. Baseball. 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

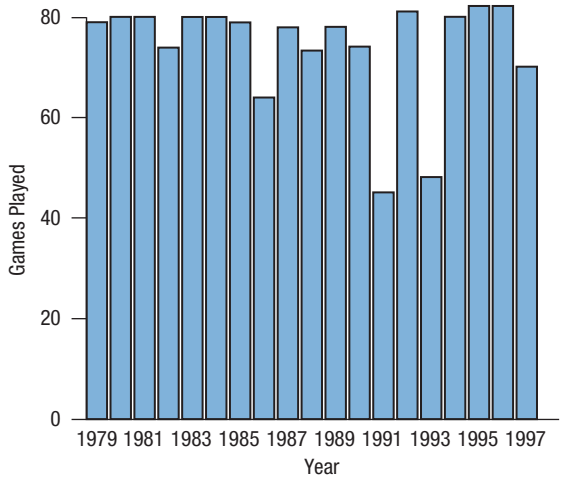
a) Create a stem-and-leaf display.
b) Sketch a boxplot.
c) Briefly describe this distribution.
d) What unusual features do you see in this distribution? What might explain this? **L.O. ❶, ❷, ❸, ❹**

40. Gretzky returns. Look once more at data of hockey games played each season by Wayne Gretzky, seen in Exercise 38.

a) Would you use the mean or the median to summarize the centre of this distribution? Why?

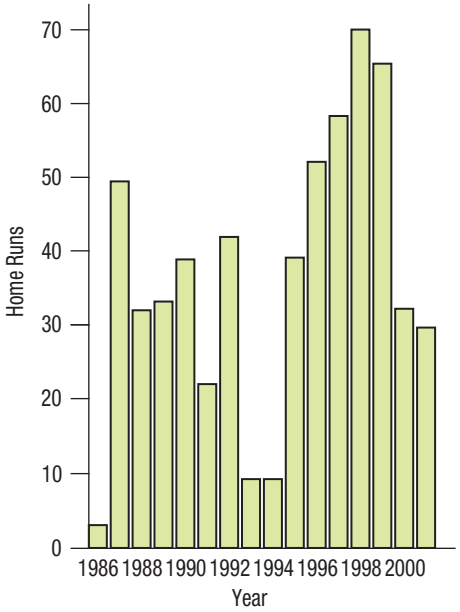
b) Without actually finding the mean, would you expect it to be lower or higher than the median? Explain.

c) A student was asked to make a histogram of the data in Exercise 38 and produced the following. Comment. **L.O. ❶, ❷**



41. McGwire, again. Look once more at data of home runs hit by Mark McGwire during his 16-year career, as seen in Exercise 39.

a) Would you use the mean or the median to summarize the centre 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 39 and produced the following. Comment. **L.O. ❶, ❷, ❹**



42. Pizza prices. The weekly prices of one brand of frozen pizza over a three-year period in Vancouver 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. **L.O. ❷, ❸, ❹, ❺**

43. Pizza prices, part 2. The weekly prices of one brand of frozen pizza over a three-year period in Montreal 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 (centre and spread) of this distribution.
e) Describe any unusual observations. **L.O. ❷, ❸, ❹, ❺**

44. Earnings of Canadians. Statistics Canada's analysis of the 2006 Census results, for Canadians employed on a full-time basis for a full year, indicates that 2.2% earned $150,000 or more in 2005 and that the median earnings grew from $40,443 in 2000 to $41,401 in 2005, a growth

rate of 0.47% per annum. Why does Statistics Canada report the median earnings instead of the mean or mode of the earnings? What would be the appropriate measure for the spread of this earnings distribution? **L.O. ❷, ❸**

45. Canadian Consumer Price Index. Calculate the mean, standard deviation, and coefficient of variation of the Consumer Price Indexes of selected Canadian cities in 2007, from the data in the following table. **L.O. ❷, ❸**

City	CPI (2007)	City	CPI (2007)
St. John's	110.7	Thunder Bay	110.8
Charlottetown	113.2	Winnipeg	111.7
and Summerside		Regina	112.7
Halifax	112	Saskatoon	117.4
Saint John	111.2	Edmonton	118
Quebec	110.1	Calgary	110.2
Montreal	110.3	Vancouver	109.8
Ottawa	110.7	Victoria	109.5
Gatineau	110.5	Whitehorse	110.8
Toronto	108.1		

Source: Statistics Canada. (2012). CANSIM Table 326-0021. Consumer Price Index (CPI), 2005 basket, annual. Retrieved from www40.statcan.ca/l01/cst01/econ45a-eng.htm

46. Canadian weekly earnings. Canadian average weekly earnings (in $) classified by province and territory are given in the table below for 2007.

a) Calculate the mean earnings for the year 2007.
b) Calculate the standard deviation for the year 2007.
c) Calculate the coefficient of variation for 2007.
d) Calculate the z-scores for Ontario and Nunavut and interpret their meaning. **L.O. ❷, ❸, ❻**

Provincial Average Weekly	Earnings in 2007
Newfoundland and Labrador	714.65
Prince Edward Island	628.90
Nova Scotia	673.38
New Brunswick	707.93
Quebec	725.29
Ontario	803.46
Manitoba	701.93
Saskatchewan	724.03
Alberta	835.52
British Columbia	761.01
Yukon	882.47
Northwest Territories	1004.63
Nunavut	948.68

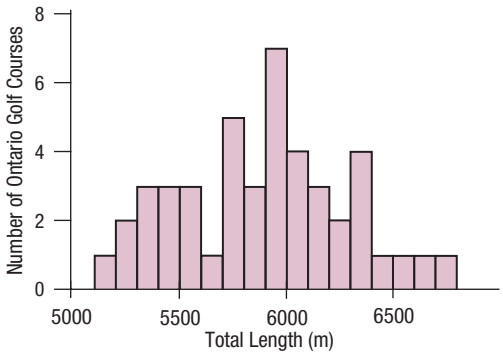
Source: Based on Statistics Canada. (2011). CANSIM Table 281-0044 and Catalogue No. 72-002-X. Earnings, average weekly, by province and territory.

47. GDP growth. Established in Paris in 1961, the Organisation for Economic Co-operation and Development (OECD) (www.oecd.org) collects information on many economic and social aspects of countries around the world, including GDP (gross domestic product). Here are the GDP growth rates of 30 industrialized countries in 2005. Write a brief report on the GDP growth rates of these countries in 2005, making sure to include appropriate graphical displays and summary statistics. **L.O. ❶, ❷, ❸, ❹, ❺**

Country	GDP 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
Netherlands	0.015
France	0.012
Germany	0.009
Portugal	0.004
Italy	0.000

48. Startup. A startup 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 Ontario. One measure of the difficulty of a golf course is its length: the total distance

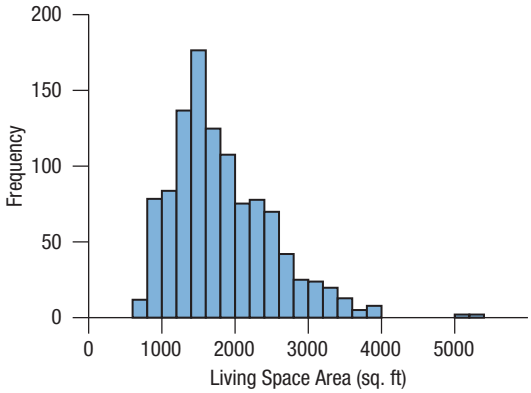
(in metres) from tee to hole for all 18 holes. Here are the histogram and summary statistics for the lengths of all the golf courses in Ontario.



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

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, centre, and spread). **L.O. ❷, ❸**

49. Real estate. A real estate agent has surveyed houses in 20 nearby postal codes in an attempt to put together a comparison for a new property she'd 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 are a histogram and summary statistics for the sizes of all the houses in the area.



Count	1057
Mean	1819.498 sq. ft
StdDev	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, centre, and spread). **L.O. ❷, ❸**

50. Food sales. Sales (in $) for one week were collected for 18 stores in a food store chain in Atlantic Canada. The stores and the towns in which the chain is located 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 the 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? **L.O. ❶, ❷, ❸, ❹, ❺**

51. Insurance profits. Life insurance companies don't know whether a policy is profitable until the policy matures (expires). To see how one company has performed recently, an analyst looked at mature policies and investigated the net profit to the company (in $).

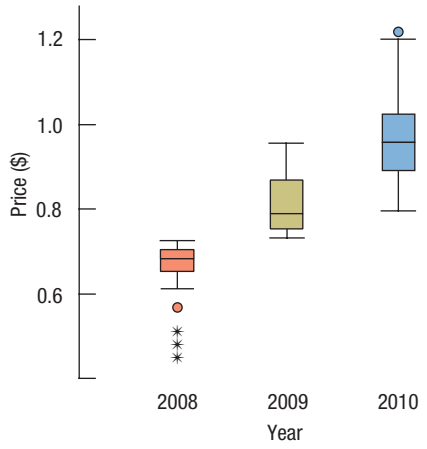
a) Make a suitable display of the profits from the data provided.
b) Summarize the central value for the profits with a median and mean. Why do they differ?
c) Given what you know about the distribution, which of these measures might do a better job of summarizing the company's profits? Why?
d) Summarize 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? **L.O. ❶, ❷, ❸, ❹, ❺**

52. 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 in the file, 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. **L.O. ❶, ❷, ❸**

53. Unemployment. The data set provided contains unemployment rates in 2008 for 23 developed countries (www.oecd.org). Produce an appropriate graphical display and briefly describe the distribution of unemployment rates. **L.O. ❶, ❷, ❸, ❹, ❺**

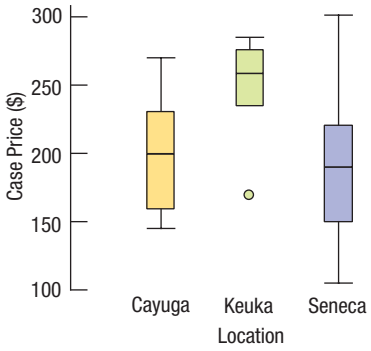
54. Gas prices, part 2. Here are boxplots of weekly gas prices at a service station in Alberta (in $/L).



a) Compare the distribution of prices over the three years.
b) In which year were the prices least stable (most volatile)? Explain. **L.O. ❷, ❸, ❹, ❺**

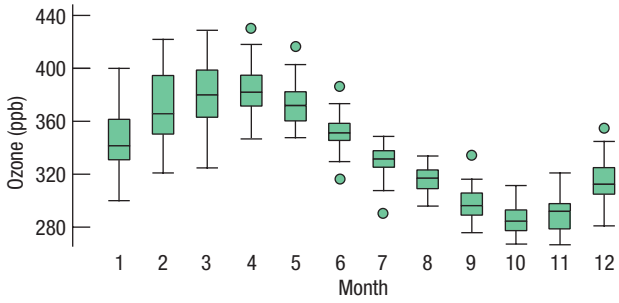
55. Fuel economy. A new hybrid car uses 3.8 litres of gasoline per 100 kilometres for city driving, according to websites advertising the car. Of course, not all of these cars are going to get the same fuel economy in all cities with all drivers. Would you expect the interquartile range (IQR) to be approximately 0.01, 1.0, or 5.0 L/100 km? Give a reason for your answer. Given your estimate of the IQR, what is your estimate of a range of reasonable values for the variance? Be sure to state the units of measurement, and give a reason for your answer. **L.O. ❷, ❸, ❹, ❺**

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



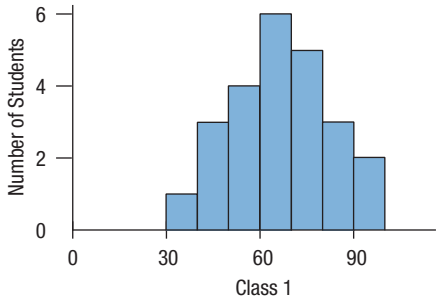
a) Which lake region produces the most expensive wine?
b) Which lake region produces the cheapest wine?
c) In which region are the wines generally more expensive?
d) Write a few sentences describing these prices. **L.O. ❷, ❸, ❹, ❺**

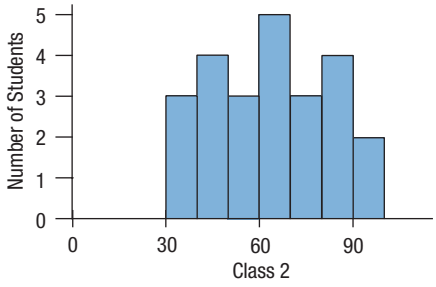
57. Ozone. Ozone levels (in parts per billion, ppb) were recorded monthly at three different sites between 1926 and 1971. Here are boxplots of the data for each month (over the 46 years), lined up in order (January = 1).

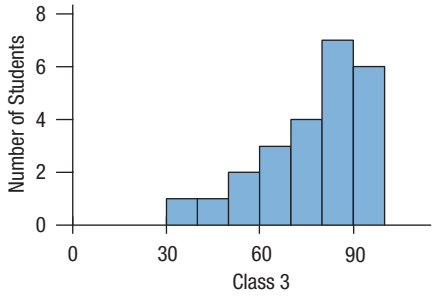


a) In what month was the highest ozone level ever recorded?
b) Which month has the largest IQR?
c) Which month has the smallest range?
d) Write a brief comparison of the ozone levels in January and June.
e) Write a report on the annual patterns you see in the ozone levels. **L.O. ❷, ❸**

58. 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? **L.O. ❷**

59. Test scores, again. Look again at the histograms of test scores for the three Statistics classes in Exercise 58.

a) Overall, which class do you think performed better on the test? Why?
b) How would you describe the shape of each distribution? **L.O. ❷, ❸, ❹**

60. Quality control. 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 centimetres) from the centre of a target on the board to the centre of the hole. The data and summary statistics are shown in the table.

	Fast	Slow
	0.000102	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 written displays of the distributions, and make a recommendation to the engineers about the accuracy of the methods. **L.O. ❷, ❸, ❹, ❺**

61. 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 recently sold homes. 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. **L.O. ❶, ❷, ❸, ❹, ❺**

62. Customer database. A philanthropic organization has a database of millions of donors whom 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, Your Imperial Majesty, and Cardinal, to name three 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*.

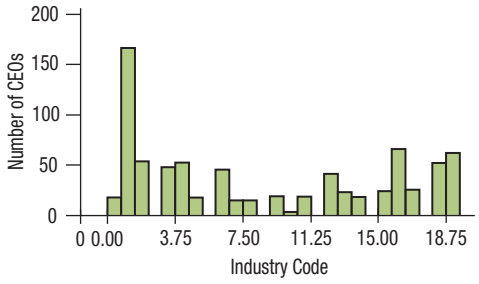
Mean	54.41
StdDev	957.62
Median	1
IQR	2
n	94649

a) What does the mean of 54.41 mean?
b) What are the typical reasons that cause measures of centre and spread to be as different as those in this table?
c) Is that why these are so different?
d) What is the basic mistake the intern made? **L.O. ❷, ❸**

63. CEOs. For each CEO, a code is listed that corresponds to the industry of the CEO's company. Here are a few 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 Nondurable	18
Forest products	9	Electric Utilities	19
Aerospace/defence	11		

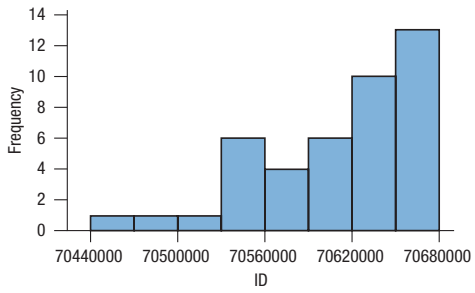
A recently hired investment analyst has been assigned to examine the industries and the number of CEOs. To start the analysis, he produces the following histogram of industry codes.



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? **L.O. ❶**

64. 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 produced 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? **L.O. ❸**

65. Car discounts, part 3. The discounts negotiated by the car buyers in Exercise 35 are classified by whether the buyer was male (code = 0) or female (code = 1). Compare the discounts of men and of women using an appropriate display and write a brief summary of the differences. **L.O. ❶, ❷, ❸, ❹**

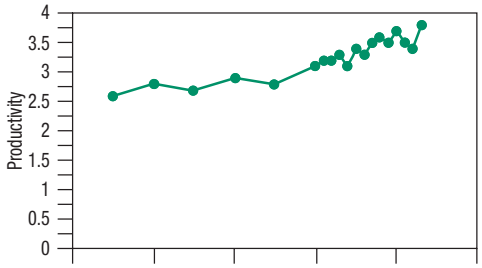
66. Hurricanes. Buying insurance for property loss from hurricanes has become increasingly difficult since Hurricane Katrina caused record property damage and loss. Many companies have refused to renew policies or write new ones. The data set provided contains the total number of hurricanes by every full decade from 1851 to 2000 (from the U.S. National Hurricane Center). Some scientists claim that the number of hurricanes has increased 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? **L.O.** 1, 7

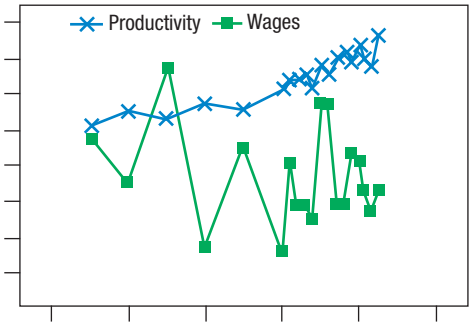
67. Hurricanes, part 2. Using the hurricanes data set, examine the number of major hurricanes (Category 3, 4, or 5) by every full decade from 1851 to 2000.

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 scientists' claim that the number of major hurricanes has been increasing (at least up to the year 2000)? **L.O. ❶, ❼**

68. Productivity study. A national productivity research centre releases information on the efficiency of workers. In a recent report, it included the following graph showing a rapid rise in productivity. What questions do you have about this? LO ❼

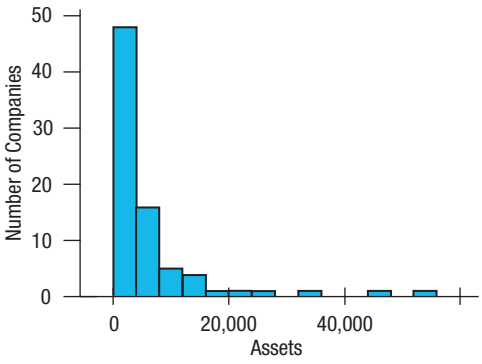


69. Productivity study revisited. A second report by the research centre analyzed the relationship between productivity and wages. Comment on the graph it used. **L.O. ❼**

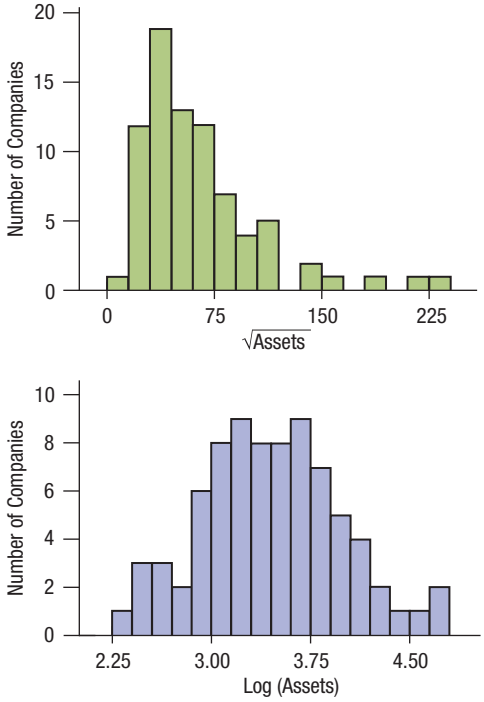


70. Assets. Here is a histogram of the assets (in millions of dollars) of 79 companies chosen from the *Forbes* list of the top U.S. corporations.

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



71. Assets, again. Here are the same data you saw in Exercise 70 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 companies' assets? **L.O. ❶**

72. Real estate, part 2. The 1057 houses described in Exercise 49 have a mean price of $167,900, with a standard deviation of $77,158. The mean living area is 1819 square feet, with a standard deviation of 663 square feet. Which is more unusual, a house in that market that sells for $400,000 or a house that has 4000 square feet of living area? Explain. **L.O. ❻**

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

	Procedures (no.)	Time (days)	Cost (%)
Mean	7.9	27.9	14.2
SD	2.9	19.6	12.9

Here are the data for three countries.

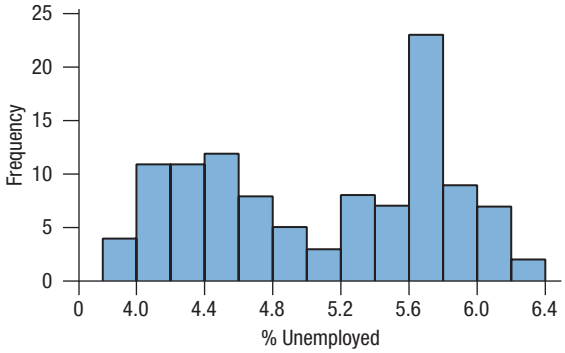
Procedures (no.)		Time	Cost
Spain	10	47	15.1
Guatemala	11	26	47.3
Fiji	8	46	25.3

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 a business. **L.O. ❻**

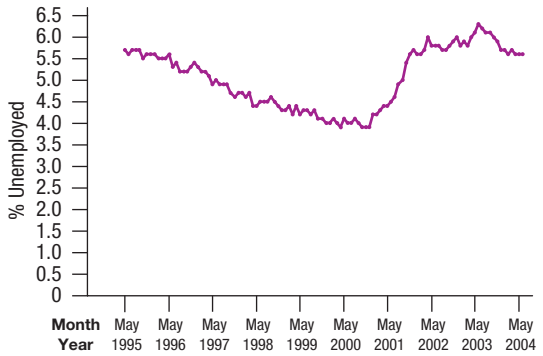
74. GDP per capita. The GDP per capita for 2007 in selected eurozone countries is given in the table. Calculate the mean, median, and standard deviation of this data. **L.O. ❻, ❸**

Austria	29,188
Cyprus	16,133
France	26,326
Germany	27,215
Greece	16,433
Ireland	41,662
Luxembourg	61,609
Malta	10,842
Portugal	12,413
Slovenia	12,983

75. Unemployment rate. The histogram shows the monthly U.S. unemployment rate from June 1995 to June 2004.



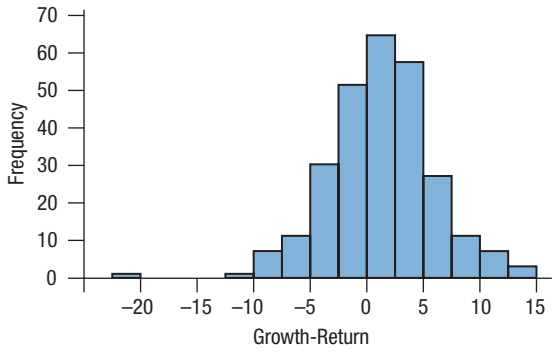
Here is the time series plot for the same data.



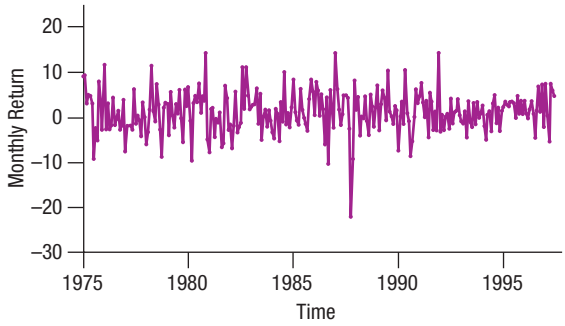
a) What features of the data can you see in the histogram that aren't clear in the time series plot?
b) What features of the data can you see in the time series plot that aren't clear in the histogram?

c) Which graphical display seems more appropriate for these data? Explain.
d) Write a brief description of unemployment rates over this time period in the United States. **L.O. ❶, ❼**

75. 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 more appropriate for these data? Explain.
d) Write a brief description of monthly mutual fund returns over this time period. **L.O. ❶, ❼**

76. Ottawa gas prices again. The actual prices at the 17 gas stations referred to in Exercise 33 were:

83.9	85.1	93.9	89.9
84.4	85.1	93.9	89.7
84.8	85.1	93.9	
84.8	88.7	93.6	
85.1	99.9	93.4	

Construct a stem-and-leaf diagram of this data and point out any unusual features that are not evident from the histogram in Exercise 33. LO ❶

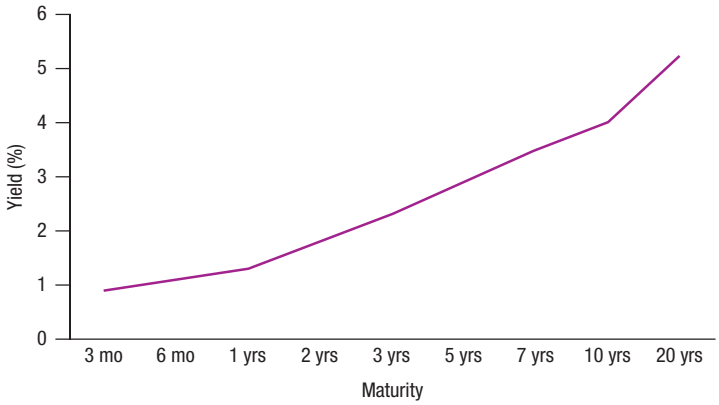
77. Food consumption. FAOSTAT, the Food and Agriculture Organization of the United Nations, collects information on the production and consumption of more than 200 food and agricultural products for 200 countries around the world. The following table lists meat consumption (per capita in kilograms per year) and alcohol consumption (per capita in gallons per year) for selected countries. The United States leads in meat consumption with 267.30 kilograms, while Ireland is the largest alcohol consumer at 55.80 gallons.

Calculate the z-scores for meat and alcohol consumption in Ireland and the United States, and interpret the meaning of the scores. **L.O. ❷, ❸, ❻**

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

78. Investments. Four people each invest $1000, with each person garnering a different rate of return.
a) The first three people invest $1000 each for one year in three different investments. The first person gets a return of 16% and the other two get 1% and 27%, respectively. What is the average return on the three investments?
b) The fourth investor invests $1000 for three years. At the end of each year he reinvests his return plus capital for the next year. He makes 16%, 1%, and 27% in the three years, respectively. What is his average rate of return over the three years? **L.O. ❷**

79. Canadian bond yields and ethics. Alfredo Wagar, an analyst, produced the graph below showing how Canadian government bond yields depend on the amount of time left until the maturity of the bond. He recommends "buying bonds with 3 month, 6 month and 20 year maturities, since their yields are above the general trend."



a) Comment on the ethics of Alfredo's recommendation as it relates to the ASA Ethical Guidelines in Appendix C.
b) Draw a better graph of the data and state the improvement(s) you have made.
c) Using your graph, do you agree with Alfredo's recommendation? **L.O. ❼**

Just Checking Answers

1 Incomes are skewed to the right, since there are a few high-income families and no family has a negative income, making the median the more appropriate measure of centre. The mean will be influenced by the high end of family incomes and will 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 9 litres per 100 kilometres would mean that only 50% of the cars get fuel efficiency in an interval 9 litres per 100 kilometres wide. Fuel economy doesn't vary that much—2 litres per 100 kilometres is reasonable. It seems plausible that 50% of the cars will be within the range 8–10 litres per 100 kilometres. An IQR of 0.1 litre per 100 kilometres would mean that the fuel efficiency of half the cars varies very little from the estimate. It's unlikely that cars, drivers, and driving conditions are that consistent.

3 We'd prefer a standard deviation of two months. Making a consistent product is important for quality. Customers want to be able to count on the MP3 player lasting somewhere close to five years, and a standard deviation of two years would mean that lifespans of the product were highly variable.