CHAPTER 3

Descriptive Statistics

Chapter Outline

> *And that's the news from Lake Wobegon, where all the women are strong, all the men are good-looking, and all the children are above-average.*
>
> ***—Garrison Keillor***

Chapter 2 explained how graphs help us understand data by identifying typical values, outliers, variations, trends, and correlations. For formal statistical inference, we need to work with precise numerical measures. It is not enough to say "It looks like output and unemployment are inversely related." We want to measure the closeness of the relationship and the size of the relationship by making statements like this: "There is a −0.85 correlation between the change in real GDP and the change in the unemployment rate; when real GDP increases by 1 percent, the unemployment rate tends to fall by about 0.4 percentage points." We would also like to have some numerical measure of how much confidence we have in such statements.

In this chapter, we look at several statistical measures used to describe data and draw statistical inferences.

3.1 Mean

The most well-known descriptive statistic is the *mean*, or average value, which is obtained by adding up the values of the data and dividing by the number of observations:

$$\begin{aligned}\overline{X} &= \frac{X_1 + X_2 + \cdots + X_n}{n} \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i\end{aligned} \tag{3.1}$$

where the n observations are denoted as X_i and the upper case Greek letter Σ (pronounced "sigma") indicates that the values should be summed up. The standard symbol for the mean, $\overline{X}$, can be pronounced "X-bar."

For example, Table 2.3 (in Chapter 2) gives the percentage changes in the prices of the 30 Dow stocks on January 29, 2008. The average is obtained by adding up the 30 values and dividing by 30:

$$\begin{aligned}\overline{X} &= \frac{3.78+3.36+\cdots-1.24}{30}\\ &= \frac{14.32}{30}\\ &= 0.48\end{aligned}$$

The average percentage change was 0.48 percent.

One interesting property of the mean is that, if we calculate how far each value is from the mean, the average of these deviations is 0:

$$\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X}) = 0$$

Table 3.1 gives three simple numerical examples. In the first example, the data are symmetrical about the mean of 20. The first observation is 10 below the mean, the second is equal to the mean, and the third is 10 above the mean. In the second example, the increase in the value of the third observation from 30 to 90 pulls the mean up to 40. Now, the first observation is 30 below the mean, the second is 20 below the mean, and the third is 50 above the mean. In the third example, the increase in the third observation to 270 pulls the mean up to 100. Now, the first observation is 90 below the mean, the second is 80 below the mean, and the third is 170 above the mean. In every case, the average deviation is 0.

Outliers are values that are very different from the other observations and pull the mean toward them. A national magazine once reported [1] that a group of Colorado teachers had failed a history test, with an average score of 67. It turned out that only four teachers had taken the test and one received a score of 20. The other three averaged 83. The one very low score pulled the mean down to 67 and misled a magazine that interpreted the average score as the typical score.

Table 3.1: The Average Deviation from the Mean Is 0

	Symmetrical Data		Data Skewed Right		Big Outlier	
	X_i	$X_i-\overline{X}$	X_i	$X_i-\overline{X}$	X_i	$X_i-\overline{X}$
	10	−10	10	−30	10	−90
	20	0	20	−20	20	−80
	30	10	90	50	270	170
Sum	60	0	120	0	300	0
Average	20	0	40	0	100	0

The Joint Economic Committee of Congress once reported that the share of the nation's wealth owned by the richest 0.5 percent of U.S. families had increased from 25 percent in 1963 to 35 percent in 1983 [2]. Politicians made speeches and newspapers reported the story with "Rich Get Richer" headlines. Some skeptics in Washington rechecked the calculations and discovered that the reported increase was due almost entirely to the incorrect recording of one family's wealth as $200,000,000 rather than $2,000,000, an error that raised the average wealth of the rich people who had been surveyed by nearly 50 percent. Someone typed two extra zeros and temporarily misled the nation.

One way to make the mean less sensitive to outliers is to discard the extreme observations. More than a hundred years ago, a renowned statistician, Francis Edgeworth, argued that the mean is improved "when we have thrown overboard a certain portion of our data—a sort of sacrifice which has often to be made by those who sail upon the stormy seas of Probability" [3]. A *trimmed mean* is calculated by discarding a specified percentage of the data at the two extremes and calculating the average of the remaining data. For example, a 10-percent trimmed mean discards the highest 10 percent of the observations and the lowest 10 percent, then calculates the average of the remaining 80 percent. Trimmed means are commonly used in many sports competitions (such as figure skating) to protect the performers from judges who are excessively generous with their favorites and harsh with other competitors.

3.2 Median

The median is the middle value of the data when the data are arranged in numerical order. It does not matter whether we count in from highest to lowest or lowest to highest. If, for example, there are 11 observations, the median is the sixth observation; if there are 10 observations, the median is halfway between the fifth and sixth observations.

In general, at least half of the values are greater than or equal to the median and at least half are smaller than or equal to the median. We have to use the qualifiers *at least* and *or equal to* because more than one observation may be exactly equal to the median, as with these data:

$$1 \qquad 2 \qquad 2 \qquad 2 \qquad 7$$

It is not true that half of the observations are greater than 2 and half are less than 2, but it is true that at least half the observations are greater than or equal to 2 and at least half are less than or equal to 2.

The Dow percentage change data in Table 2.1 are arranged in numerical order. Because there are 30 stocks, the median is halfway between the 15th and 16th returns (Citigroup 0.26 and Wal-Mart 0.30):

$$\begin{aligned}\text{median} &= \frac{0.26 + 0.30}{2}\\ &= 0.28\end{aligned}$$

Earlier, we calculated the mean of these data to be 0.48. A histogram (Figure 2.10) of the Dow percentage changes shows that the data are roughly symmetrical, but the two large values (Boeing 3.36 and Alcoa 3.78) pull the mean somewhat above the median.

In comparison to the mean, the median is more robust or resistant to outliers. Looking again at the data in Table 3.1, the median stays at 20 while the increase in the value of the third observation from 30 to 90 to 270 pulls the mean upward.

For some questions, the mean is the most appropriate answer. Whether you balance your budget over the course of a year depends on your average monthly income and expenses. Farm production depends on the average yield per acre. The total amount of cereal needed to fill a million boxes depends on the average net weight. For other questions, the median might be more appropriate.

The U.S. Census Bureau reports both the mean and median household income. The mean income tells us how much each person would earn if the total income were equally divided among households. Of course, income is not evenly divided. In 2007, the mean household income in the United States was $69,193 while the median was $50,740. The mean was pulled above the median by a relatively small number of people with relatively high incomes.

3.3 Standard Deviation

An average does not tell us the underlying variation in the data. Sir Francis Galton once commented that “It is difficult to understand why statisticians commonly limit their enquiries to Averages, and do not revel in more comprehensive views. Their souls seem as dull to the charm of variety as that of the native of one of our flat English counties, whose retrospect of Switzerland was that, if its mountains could be thrown into its lakes, two nuisances would be got rid of at once” [4]. Another Englishman with a sense of humor, Lord Justice Matthews, once told a group of lawyers: “When I was a young man practicing at the bar, I lost a great many cases I should have won. As I got along, I won a great many cases I ought to have lost; so on the whole, justice was done” [5].

We might try to measure the variation or spread in the data by calculating the average deviation from the mean. But remember that, as in Table 3.1, the average deviation from the mean is always 0, no matter what the data look like. Because the positive and negative deviations from the mean offset each other, giving an average deviation of 0, we learn nothing at all about the sizes of the deviations. To eliminate this offsetting of positive and negative deviations, we could take the absolute value of each deviation and then calculate the *average absolute deviation*, which is the sum of the absolute values of the deviations from the mean, divided by the number of observations.

The average absolute deviation is easy to calculate and interpret; however, there are two reasons why it is seldom used. First, while it is easily calculated, a theoretical analysis is very difficult. Second, there is an attractive alternative that plays a prominent role in probability and statistics.

Remember that we took absolute values of the deviations to keep the positive and negative deviations from offsetting each other. Another technique that accomplishes this same end is to square each deviation, since the squares of positive and negative deviations are both positive. The average of these squared deviations is the *variance* s^2 and the square root of the variance is the *standard deviation* s:

$$s^2 = \frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}{n-1} \tag{3.2}$$

Notice that the variance of a set of data is calculated by dividing the sum of the squared deviations by $n - 1$, rather than n. In later chapters, we look at data that are randomly selected from a large population. It can be shown mathematically that, if the variance in the randomly selected data is used to estimate the variance of the population from which these data came, this estimate will, on average, be too low if we divide by n, but will, on average, be correct if we divide by $n - 1$. (This is an example of the kinds of general theorems that mathematicians can prove when they work with the standard deviation rather than the average absolute deviation.)

The variance and standard deviation are equivalent measures of the dispersion in a set of data, in that a data set that has a higher variance than another data set also has a higher standard deviation. However, because each deviation is squared, the variance has a scale that is much larger than the underlying data. The standard deviation has the same units and scale as the original data.

3.4 Boxplots

After the data are arranged in numerical order, from smallest to largest, the data can be divided into four groups of equal size, known as quartiles:

First quartile	minimum to 25th percentile
Second quartile	25th percentile to 50th percentile
Third quartile	50th percentile to 75th percentile
Fourth quartile	75th percentile to maximum

These four subsets are called *quartiles* because each encompasses one-fourth of the data. The upper limit of the first quartile is the 25th percentile because this value is larger than 25 percent of the data. Similarly, the upper limit of the second quartile is the 50th percentile, and the upper limit of the third quartile is the 75th percentile.

The 50th percentile is the median, and we have already seen that this is an appealing measure of the center of the data. An appealing measure of the spread of the data is the *interquartile range*, which is equal to the difference between the 25th and 75th percentiles. The interquartile range is (roughly) the range encompassed by the middle half of the data.

A boxplot (also called, more descriptively, a *box-and-whisker diagram*) uses five statistics to summarize the center and spread of the data: the smallest observation, the 25th percentile, the 50th percentile, the 75th percentile, and the largest observation. A boxplot may also identify any observations that are outliers.

Figure 3.1 shows a boxplot using the Dow daily price change data in Table 2.3. A box is used to connect the 25th and 75th percentiles. Because the ends of the box are at the 25th and 75th percentiles, the width of the box is equal to the interquartile range and encompasses the middle half of the data. The median is denoted by the line inside the box. The ends of the two horizontal lines coming out of the box (the "whiskers") show the minimum and maximum values. A boxplot conveys a considerable amount of information, but is less complicated than the histogram in Figure 2.9. It is also relatively robust, in that the box itself (but not the whiskers) is resistant to outliers. A boxplot can be drawn either horizontally or vertically.

In a modified boxplot, outliers that are farther than 1.5 times the interquartile range from the box are shown as separate points, and the whiskers stop at the most extreme points that are not outliers. Figure 3.2 is a modified boxplot for the Dow price changes in Figure 3.1, with the Boeing 3.36 percent and Alcoa 3.78 percent outliers identified.

In practice, the data will usually not divide perfectly into four groups of exactly equal size, and some statisticians (and statistics software) use slightly different rules for calculating the first and third quartiles.

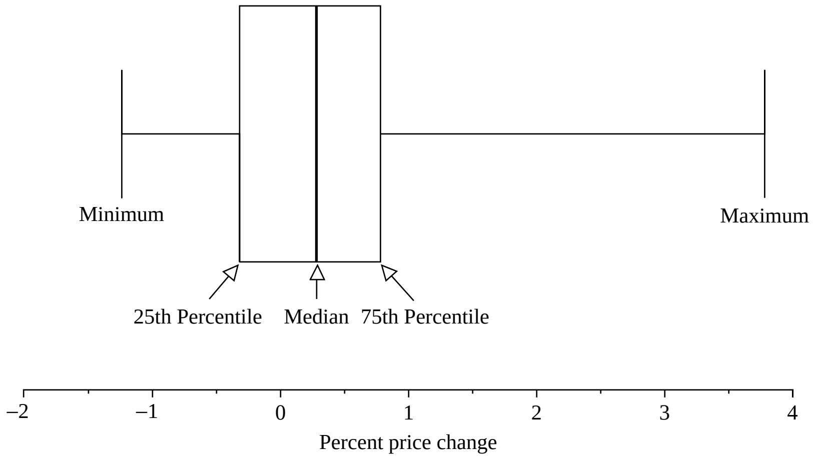

Figure 3.1
A boxplot for the Dow price changes.

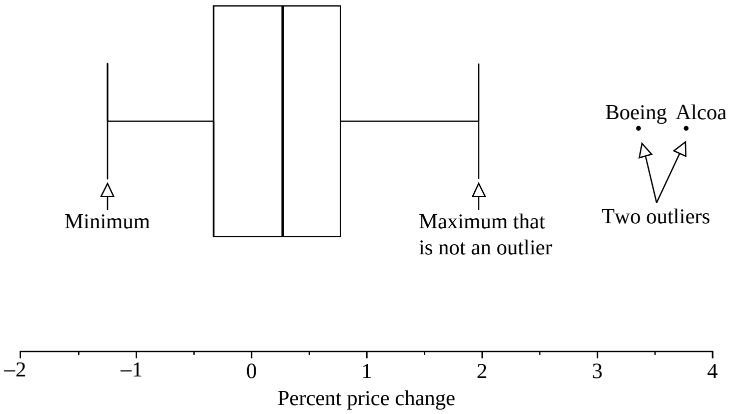

Figure 3.2
A boxplot for the Dow price changes, with two outliers.

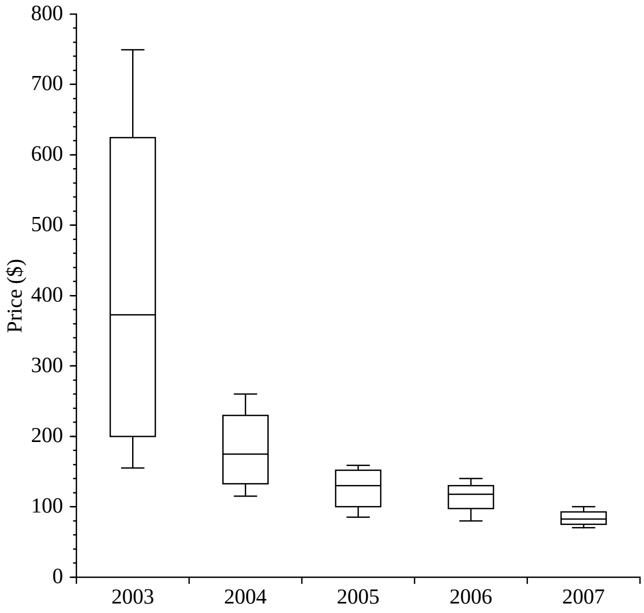

Figure 3.3
Side-by-side boxplots for hard drive prices.

Side-by-side boxplots allow us to compare data sets. Figure 3.3 shows side-by-side boxplots for the hard drive prices in Table 1.6. This figure clearly shows that, not only was there a decline in prices between 2003 and 2007, but the dispersion in prices also declined each year.

In most cases, the standard deviation, the average absolute deviation, and the interquartile range agree in their rankings of the amount of variation in different data sets. Exceptions can occur with outliers because the squaring of deviations can have a large effect on the standard deviation. Just as the median is more resistant to outliers than the mean, so the average absolute deviation and the interquartile range are more resistant to outliers than the standard deviation. Nonetheless, because of its mathematical tractability and importance in probability and statistics, the standard deviation is usually used to gauge the variation in a set of data.

3.5 Growth Rates

Many data are more meaningful when expressed as percentages; for example, we can readily interpret statements such as the CPI increased by 3 percent last year, the Dow went down 2 percent yesterday, or it costs 20 percent more to rent a home than to buy one.

Percentages can also put things in proper perspective. If the Dow falls by 340, is that a lot or a little? When the Dow was at 14,000 in 2007, a 340-point drop was a 2.4 percent decline—sickening but not catastrophic. When the Dow was at 381 in 1929, the 340-point drop that occurred between 1929 and 1932 was a catastrophic 89 percent decline.

In general, the percentage change is computed by dividing the difference between the new and old values by the value *before* the change, then multiplying by 100 to convert a fraction to a percentage:

$$\text{percentage change} = 100\left(\frac{\text{new} - \text{old}}{\text{old}}\right) \tag{3.3}$$

Get the Base Period Right

Percentage changes are sometimes calculated incorrectly by dividing the change by the value *after* the change instead of dividing by the value *before* the change. Thus, when income doubles from \$60,000 to \$120,000, the percentage change might be incorrectly reported as 50 percent,

$$\begin{aligned}\text{incorrect percentage change} &= 100\left(\frac{\text{new} - \text{old}}{\text{new}}\right)\\ &= 100\left(\frac{\$120{,}000 - \$60{,}000}{\$120{,}000}\right)\\ &= 50\end{aligned}$$

instead of 100 percent

$$\begin{aligned}\text{correct percentage change} &= 100\left(\frac{\text{new} - \text{old}}{\text{old}}\right)\\ &= 100\left(\frac{\$120{,}000 - \$60{,}000}{\$60{,}000}\right)\\ &= 100\end{aligned}$$

Newsweek once reported that the salaries of some Chinese government officials had been reduced 300 percent [6]. Suppose that the salary before the reduction were \$60,000. If the salary were eliminated completely, that would be a 100 percent reduction:

$$
\begin{aligned}
\text{percentage change} &= 100\left(\frac{\text{new} - \text{old}}{\text{old}}\right)\\
&= 100\left(\frac{\$0 - \$60{,}000}{\$60{,}000}\right)\\
&= -100
\end{aligned}
$$

To get a 300 percent reduction, the salary would have to be *negative* \$120,000!

$$
\begin{aligned}
\text{percentage change} &= 100\left(\frac{\text{new} - \text{old}}{\text{old}}\right)\\
&= 100\left(\frac{-\$120{,}000 - \$60{,}000}{\$60{,}000}\right)\\
&= -300
\end{aligned}
$$

Since it is unlikely that anyone would pay \$120,000 a year to work for the government, *Newsweek* must have made a mistake when it calculated the percentage change. In fact, it made several mistakes by calculating the percentage change this way:

$$
\begin{aligned}
\textit{Newsweek}\text{ ``percentage change''} &= -100\left(\frac{\text{old}}{\text{new}}\right)\\
&= -100\left(\frac{\$60{,}000}{\$20{,}000}\right)\\
&= -300
\end{aligned}
$$

The negative sign appears because the salary was reduced.

To calculate the percentage change correctly, we divide the change in the salary by the old salary, as in Equation 3.3. If a \$60,000 salary is reduced to \$20,000, this is not a 300 percent decrease, but rather a 66.67 percent decrease:

$$
\begin{aligned}
\text{percentage change} &= 100\left(\frac{\text{new} - \text{old}}{\text{old}}\right)\\
&= 100\left(\frac{\$20{,}000 - \$60{,}000}{\$60{,}000}\right)\\
&= -66.67
\end{aligned}
$$

For a percentage calculation to be informative, it should also use a reasonable base period that is clearly identified. If someone reports that "The price of orange juice increased 50 percent," he or she should tell us the base period. Did the price of orange juice increase 50 percent in a single day or over the past 100 years? If someone does not specify the base period (or gives a peculiar base period), that person may be trying to magnify or minimize the percentage change to support a position, rather than present the facts honestly. To make the price increase seem large, someone may compare it to a period long ago when orange juice was cheap. Someone else, who wants to report a small price increase, may choose a base period when orange juice was expensive, perhaps after bad weather ruined most of the crop. An honest statistician compares the price to a natural base period and identifies that base period: "The price of orange juice has increased by 2 percent during the past year, and by 35 percent over the past ten years." If orange juice prices fluctuate considerably, we can use a time series graph, as explained in Chapter 2, to show these ups and downs.

Because there usually is no room to specify the base period in a brief newspaper headline, we sometimes see accurate, but seemingly contradictory, headlines on the very same day. For example, the *New York Times* once reported that "American Express Net Climbs 16%" [7], while *The Wall Street Journal* reported the same day that "Net at American Express Fell 16% in 4th Quarter" [8]. These were not typographical errors. Both stories were accurate and each explained the base period that they used. The *Times* compared American Express's earnings in the fourth quarter to its earnings in the third quarter; the *Journal* compared American Express's earnings in the fourth quarter to its earnings a year earlier.

Watch out for Small Bases

Because a percentage is calculated relative to a base, a very small base can give a misleadingly large percentage. On your second birthday, your age increased by 100 percent. When you graduate from college, your income may increase by several thousand percent. On July 25, 1946, the average hourly rainfall between 6 A.M. and noon in Palo Alto, California, was 84,816 times the average hourly rainfall in July during the preceding 36 years, dating back to 1910, when the Palo Alto weather station first opened [9]. How much rain fell on this incredibly wet July 25? Only 0.19 inches. During the preceding 36 years, there had been only one July day with measurable precipitation in Palo Alto, and on that day the precipitation was 0.01 inches.

One way to guard against the small-base problem is to give the level as well as the percentage; for example, to note that the rainfall was 0.19 inches on this unusually wet July day in Palo Alto.

The Murder Capital of Massachusetts

Wellfleet is a small town on Cape Cod, renowned for its oysters, artists, and tranquility. There was considerable surprise when the Associated Press reported that Wellfleet had the highest murder rate in Massachusetts in 1993, with 40 murders per 100,000 residents—more than double the murder rate in Boston, which had only 17 murders per 100,000 residents [10]. A puzzled newspaper reporter looked into this statistical murder mystery. She found that there had in fact been no murders in Wellfleet in 1993 and that no Wellfleet police officer, including one who had lived in Wellfleet for 48 years, could remember a murder ever occurring in Wellfleet.

However, a man accused of murdering someone in Barnstable, which is 20 miles from Wellfleet, had turned himself in at the Wellfleet police station in 1993 and the Associated Press had erroneously interpreted this Wellfleet arrest as a Wellfleet murder. Because Wellfleet had only 2,491 permanent residents, this one misrecorded murder arrest translated into 40 murders per 100,000 residents. Boston, in contrast, had 98 murders, which works out to 17 murders per 100,000 residents.

The solution to this murder mystery shows how a statistical fluke can make a big difference if the base is small. A misrecorded murder in Boston would not noticeably affect its murder rate. In Wellfleet, a misrecorded murder changes the reported murder rate from 0 to 40. One way to deal with small bases is to average the data over several years in order to get a bigger base. Wellfleet's average murder rate over the past 50 years is 1 with the misrecorded arrest and 0 without it—either way confirming that it is indeed a peaceful town.

The Geometric Mean (Optional)

The arithmetic mean of n numbers is obtained by adding together the numbers and dividing by n. The geometric mean is calculated by *multiplying* the numbers and taking the nth root:

$$g = (X_1 X_2 \ldots X_n)^{1/n}, \; X_i \geq 0 \tag{3.4}$$

For example, the arithmetic mean of 5, 10, and 15 is 10:

$$\begin{aligned} \overline{X} &= \frac{5+10+15}{3} \\ &= 10 \end{aligned}$$

while the geometric mean is 9.09:

$$\begin{aligned} g &= (5 \times 10 \times 15)^{1/3} \\ &= 750^{1/3} \\ &= 9.09 \end{aligned}$$

The geometric mean is not defined for negative values and is simply equal to 0 if one of the observations is equal to 0. One interesting mathematical fact about the geometric mean is that it is equal to the arithmetic mean if all of the numbers are the same; otherwise, as in our numerical example, the geometric mean is less than the arithmetic mean.

Generally speaking, the geometric mean is pretty close to the arithmetic mean if the data are bunched together and far below the arithmetic mean if the data are far apart. For example, for the data 4, 5, 6, the arithmetic mean is 5 and the geometric mean is 4.93; for the data 10, 50, 90, the arithmetic mean is 50 and the geometric mean is 35.57.

The geometric mean is often used for data that change over time because annual rates of change are often more meaningful than the overall percentage change. For example, which of these statements (both of which are true) is more memorable and meaningful?

a. The CPI increased 154 percent between December 1980 and December 2010.
b. The CPI increased 3.2 percent a year between December 1980 and December 2010.

Statement b is based on a geometric mean calculation.

Table 3.2 shows the annual rate of inflation using December U.S. CPI data. The CPI increased by 8.9 percent from December 1980 to December 1981 and by 3.8 percent from December 1981 to December 1982. Since the level of the CPI is arbitrary, we might set the CPI equal to 1 in December 1980. The CPI is consequently 1.089 in December 1981 and (1.089)(1.038) = 1.130 in December 1982. If we continue for all 30 years, the level of the CPI in December 2010 is

$$(1.089)(1.038)(1.038)\ldots(1.015) = 2.54$$

The increase in the CPI from 1 in December 1980 to 2.54 in December 2010 is a 154 percent increase.

The geometric mean tells us that the annual rate of increase over these 30 years was about 3.2 percent:

$$[(1.089)(1.038)(1.038)\ldots(1.015)]^{1/30} = 1.032$$

Table 3.2: Annual Rate of Inflation, December to December

1981	8.9	1996	3.3
1982	3.8	1997	1.7
1983	3.8	1998	1.6
1984	3.9	1999	2.7
1985	3.8	2000	3.4
1986	1.1	2001	1.6
1987	4.4	2002	2.4
1988	4.4	2003	1.9
1989	4.6	2004	3.3
1990	6.1	2005	3.4
1991	3.1	2006	2.5
1992	2.9	2007	4.1
1993	2.7	2008	0.1
1994	2.7	2009	2.7
1995	2.5	2010	1.5

The CPI increased by an average of 3.2 percent a year in that if the CPI increased by 3.2 percent *every* year, it would be 154 percent higher after 30 years:

$$(1.032)^{30} = 2.54$$

In general, if something grows at a rate R_i in year i, then a geometric mean calculation gives us the overall annual rate of growth R:

$$1 + R = [(1 + R_1)(1 + R_2) \ldots (1 + R_n)]^{1/n} \tag{3.5}$$

Equation 3.5 also works for interest rates and other rates of return. If you invest $100 and earn a 5 percent return the first year, a 20 percent return the second year, and a 10 percent return the third year, your $100 grows to

$$\$100(1 + 0.05)(1 + 0.20)(1 + 0.10) = \$138.60$$

Using Equation 3.5, the geometric return is 11.5 percent:

$$\begin{aligned} 1 + R &= [(1 + 0.05)(1 + 0.20)(1 + 0.10)]^{1/3} \\ &= 1.386^{1/3} \\ &= 1.115 \end{aligned}$$

in that $100 invested for three years at 11.5 percent will grow to $138.60:

$$\$100(1.115)^3 = \$138.60$$

3.6 Correlation

Often, we are more interested in the relationships among variables than in a single variable in isolation. For example, how is household spending related to income? How is GDP related to the unemployment rate? How are the outcomes of presidential elections related to the unemployment rate? How are interest rates related to the rate of inflation?

Chapter 2 discusses the Federal Reserve's stock-valuation model, which hypothesizes that there is a relationship between the earnings/price ratio and the interest rate on 10-year Treasury bonds. A time series graph (Figure 2.23) suggests that these two variables do, indeed, move up and down together. Figure 3.4 shows a scatterplot of the quarterly data used in Figure 2.23. This is another way of seeing that there appears to be a positive relationship between these two variables.

The covariance s_{xy} is a very simple measure of the relationship between two variables:

$$s_{xy} = \frac{(X_1 - \overline{X})(Y_1 - \overline{Y}) + (X_2 - \overline{X})(Y_2 - \overline{Y}) + \cdots + (X_n - \overline{X})(Y_n - \overline{Y})}{n - 1} \tag{3.6}$$

The covariance is analogous to the variance, which is a measure of the dispersion in a single variable. Comparing Equations 3.2 and 3.6, we see that the variance adds up the

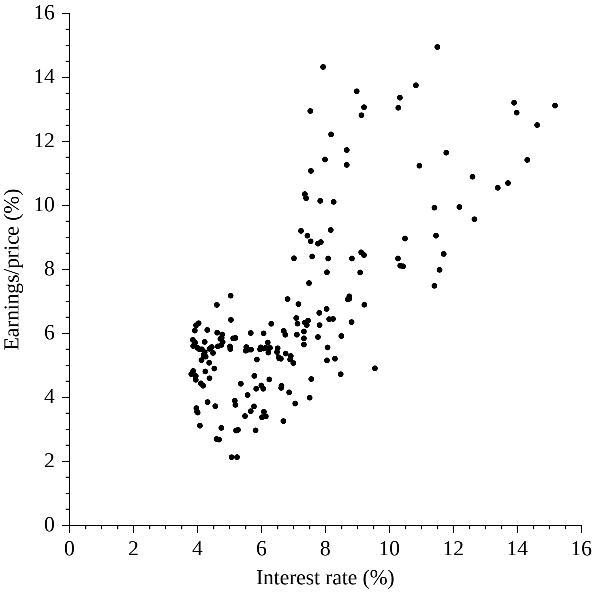

Figure 3.4
S&P 500 earnings/price ratio and 10-year Treasury interest rate, 1960–2009.

squared deviations of a single variable from its mean, while the covariance adds up the products of the deviations of two variables from their means.

Figure 3.5 repeats Figure 3.4, this time showing the mean values of the two variables, so that we can see when each of these variables is above or below its mean. Clearly, when the interest rate is above its mean, the earnings/price ratio tends to be above its mean, and when the interest rate is below its mean, the earnings/price ratio tends to be below its mean.

Now look again at the formula for the covariance. The product of the deviations of X and Y from their respective means is positive when both deviations are positive (the northeast quadrant of Figure 3.5) or when both deviations are negative (the southwest quadrant of Figure 3.5). The product of the deviations of X and Y from their respective means is negative when one deviation is positive and the other deviation is negative (the northwest and southeast quadrants of Figure 3.5). In the case of the 10-year Treasury rate and the S&P 500 earnings/price ratio, the data are overwhelmingly in the northeast and southwest quadrants, and the products of the deviations are overwhelmingly positive. So, the covariance is positive.

This reasoning shows us how to interpret a positive or negative covariance:

> **Positive covariance**: When the value of one variable is above (or below) its mean, the value of the other variable also tends to be above (or below) its mean.
> **Negative covariance**: When the value of one variable is above its mean, the value of the other variable tends to be below its mean, and vice versa.

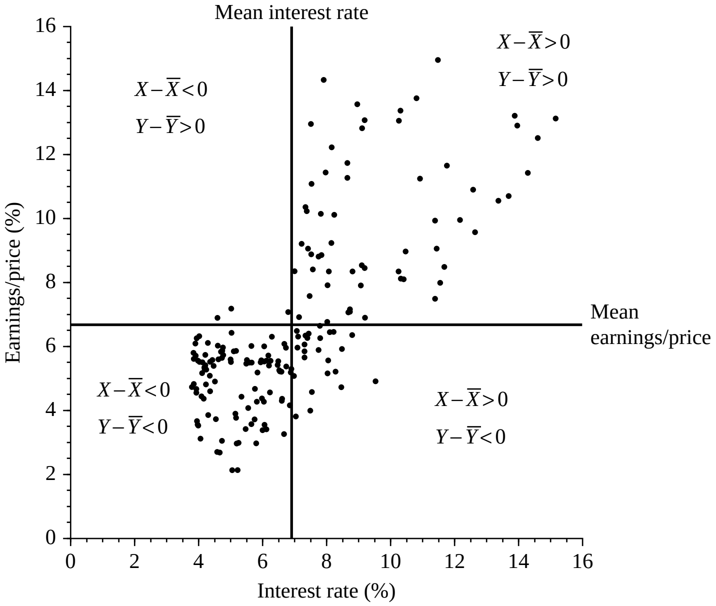

Figure 3.5
S&P 500 earnings/price ratio and 10-year Treasury rate, relative to means.

In our example, the covariance between the 10-year Treasury rate and the S&P 500 earnings/price ratio works out to be 5.14. When the 10-year Treasury rate is above (or below) its average, the S&P 500 earnings/price ratio tends to be above (or below) its average value, too.

Is a 5.14 covariance a big number or a small number? The size of the covariance depends on the scale of the deviations from the means. If the typical deviation is 100, the covariance will clearly be much higher than when the typical deviation is 1. To put the covariance into perspective, we calculate the *correlation R* by dividing the covariance by the standard deviation of *X* and the standard deviation of *Y*:

$$\begin{aligned} R &= \frac{(\text{covariance between } X \text{ and } Y)}{(\text{standard deviation of } X)(\text{standard deviation of } Y)} \\ &= \frac{s_{XY}}{s_X s_Y} \end{aligned} \tag{3.7}$$

Standard deviations are always positive. Therefore, the correlation coefficient *R* is positive when the covariance is positive and negative when the covariance is negative. A positive correlation means that, when either variable is above its mean value, the other variable tends to be above its mean value. A negative correlation means that, when either variable is above its mean value, the other variable tends to be below its mean value.

It can be shown mathematically that the value of the correlation coefficient cannot be larger than 1 or less than −1. The correlation is equal to 1 if all the data lie on a straight line with a positive slope and is equal to −1 if all the data lie on a straight line with a negative slope. In each of these extreme cases, if you know the value of one of the variables, you can be certain of the value of the other variable. In less extreme cases, we can visualize drawing a positively sloped or negatively sloped straight line through the data, but the data do not lie exactly on the line. The correlation is equal to 0 when there is no linear relationship between the two variables and we cannot visualize drawing a straight line through the data.

The value of the correlation coefficient does not depend on which variable is on the vertical axis and which is on the horizontal axis. For this reason, the correlation coefficient is often used to give a quantitative measure of the direction and degree of association between two variables that are related statistically but not necessarily causally. For instance, the historical correlation between midterm and final examination scores in my introductory statistics classes is about 0.6. I do not believe that either test score has any causal effect on the other test score but rather that both scores are influenced by common factors (such as student aptitude and effort). The correlation coefficient is a numerical measure of the degree to which these test scores are related statistically.

The correlation coefficient does not depend on the units in which the variables are measured—pennies or dollars, inches or centimeters, pounds or kilograms, thousands or millions. For instance, the correlation between height and weight does not depend on whether the heights are measured in inches or centimeters or the weights are measured in pounds or kilograms.

The correlation between the 10-year Treasury rate and the S&P 500 earnings/price ratio turns out to be 0.73, which shows (as suggested by Figures 3.4 and 3.5) that there is a positive, but certainly not perfect, linear relationship between these two variables.

Figure 3.6 shows a scatterplot of annual data on the percent change in real GDP and the change in the unemployment rate. This time, the correlation is −0.85, which shows that there is a strong negative linear relationship between these two variables.

Figure 3.7 shows a scatterplot for two essentially unrelated variables, the rate of return on the S&P 500 and per capita cigarette consumption. The correlation is 0.04, which confirms our visual impression that there is no meaningful linear relation between these two variables. If we try to visualize a straight line fit to these data, we would not know where to draw it.

A correlation of 0 does not mean that there is no relationship between X and Y, only that there is no linear relationship. Figure 3.8 shows how a zero correlation coefficient does not rule out a perfect nonlinear relationship between two variables. In this figure, Y and X are exactly related by the equation $Y = 10X - X^2$. Yet, because there is no linear relationship between Y and X, the correlation coefficient is 0. This is an example of why we should look at a scatterplot of our data.

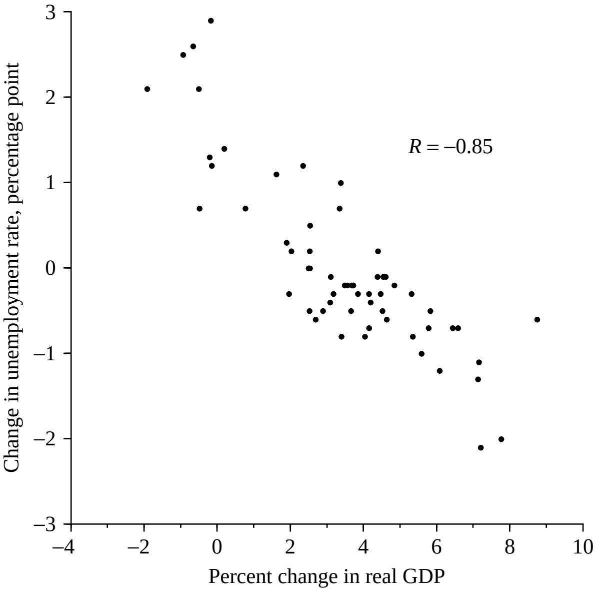

Figure 3.6
Real GDP and unemployment, annual, 1948–2008.

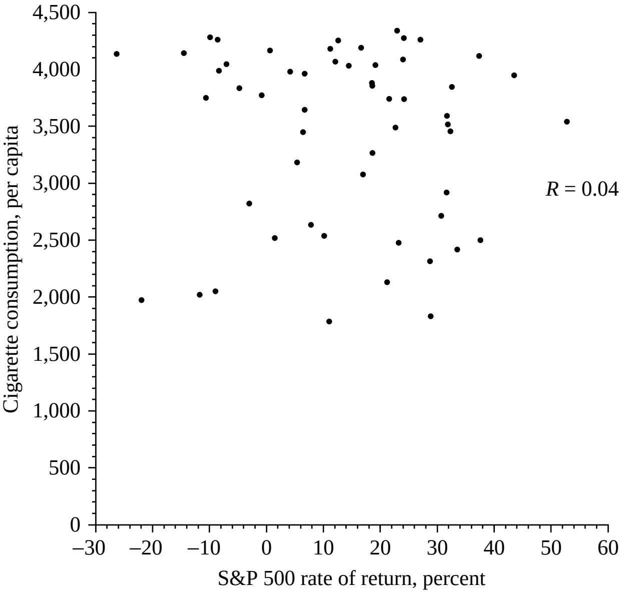

Figure 3.7
S&P 500 annual rate of return and per capita cigarette consumption, 1950–2007.

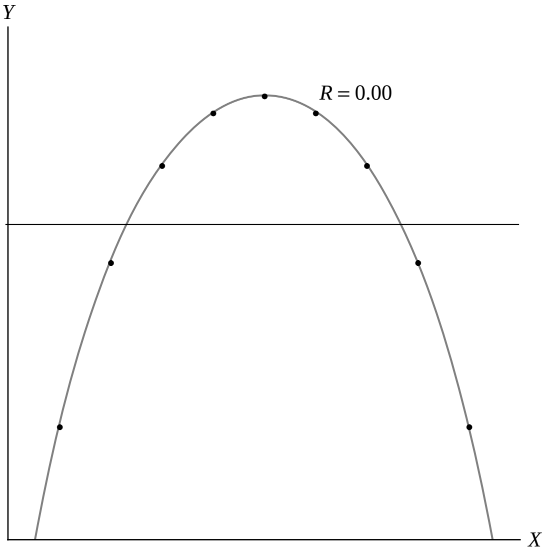

Figure 3.8
A zero correlation does not rule out a nonlinear relationship.

Exercises

3.1 Which of these data sets has a higher mean? Higher median? Higher standard deviation? (Do not do any calculations. Just look at the data.)

X	1	2	3	4	5
Y	5	4	3	2	1

3.2 Which of these data sets has a higher mean? Higher median? Higher standard deviation? (Do not do any calculations. Just look at the data.)

X	1	2	3	4	5
Y	1	2	3	4	6

3.3 Albert Michelson's 1882 measurements of the speed of light in air (in kilometers per second) were as follows [11]:

299,883	299,796	299,611	299,781	299,774	299,696
299,748	299,809	299,816	299,682	299,599	299,578
299,820	299,573	299,797	299,723	299,778	299,711
300,051	299,796	299,772	299,748	299,851	

Calculate the mean, median, and 10-percent trimmed mean. Which is closest to the value 299,710.5 that is now accepted as the speed of light?

3.4 In 1798, Henry Cavendish made 23 measurements of the density of the earth relative to the density of water [12]:

5.10	5.27	5.29	5.29	5.30	5.34	5.34	5.36	5.39	5.42	5.44	5.46
5.47	5.53	5.57	5.58	5.62	5.63	5.65	5.68	5.75	5.79	5.85	

Calculate the mean, median, and 10-percent trimmed mean. Which is closest to the value 5.517 that is now accepted as the density of the earth?

3.5 Twenty-four female college seniors were asked how many biological children they expect to have during their lifetimes. Figure 3.9 is a histogram summarizing their answers: Of these four numbers (0, 1.1, 2, 2.3), one is the mean, one is the median, one is the standard deviation, and one is irrelevant. Identify which number is which.

a. Mean.
b. Median.
c. Standard deviation.
d. Irrelevant.

3.6 Thirty-five male college seniors were asked how many biological children they expect to have during their lifetimes. Figure 3.10 is a histogram summarizing their answers: Of these four numbers (0, 1.4, 2.4, 3.0), one is the mean, one is the median, one is the standard deviation, and one is irrelevant. Identify which number is which.

a. Mean.
b. Median.
c. Standard deviation.
d. Irrelevant.

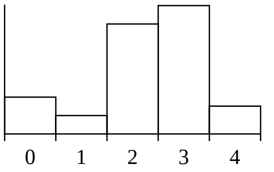

Figure 3.9
Exercise 3.5.

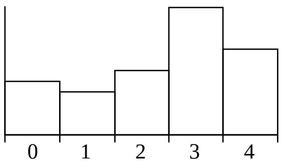

Figure 3.10
Exercise 3.6.

3.7 Suppose that you have ten observations that have a mean of 7, a median of 6, and a standard deviation of 3. If you add 5 to the value of each observation, what are the new values of the

a. Mean?
b. Median?
c. Standard deviation?

3.8 Suppose that you have ten observations that have a mean of 7, a median of 6, and a standard deviation of 3. If you subtract 2 from the value of each observation, what are the new values of the

a. Mean?
b. Median?
c. Standard deviation?

3.9 Suppose that you have ten observations that have a mean of 7, a median of 6, and a standard deviation of 3. If you double the value of each observation, what are the new values of the

a. Mean?
b. Median?
c. Standard deviation?

3.10 Suppose that you have ten observations that have a mean of 7, a median of 6, and a standard deviation of 3. If you halve the value of each observation, what are the new values of the

a. Mean?
b. Median?
c. Standard deviation?

3.11 Identify the apparent statistical mistake in this commentary [13]:

> *The median cost of a house in [Duarte, California] is a whopping $4,276,462, making it the most expensive housing market in the country. It ranks No. 1 on Forbes' annual ranking of America's Most Expensive ZIP Codes.... [O]nly 12 homes are currently on the market. So a single high-priced listing (like the mammoth nine-bedroom, built this year, that's selling for $19.8 million) is enough to skew the median price skyward.*

3.12 Roll four six-sided dice ten times, each time recording the sum of the four numbers rolled. Calculate the mean and median of your ten rolls. Repeat this experiment 20 times. Which of these two measures seems to be the least stable?

3.13 Use the test scores in Exercise 1.8 to calculate the average score for each grade level and then make a time series graph using these average scores.

3.14 An old joke is that a certain economics professor left Yale to go to Harvard and thereby improved the average quality of both departments. Is this possible?

3.15 Ann Landers, an advice columnist, once wrote "Nothing shocks me anymore, especially when I know that 50 percent of the doctors who practice medicine graduated in the bottom half of their class" [14]. Does this observation imply that half of all doctors are incompetent?

3.16 When is the standard deviation negative?

3.17 There was a players' strike in the middle of the 1981 baseball season. Each division determined its season winner by having a playoff between the winner of the first part of the season, before the strike, and the winner of the second part of the season. Is it possible for a team to have the best winning percentage for the season as a whole, yet not win either half of the season and consequently not have a chance of qualifying for the World Series? Use some hypothetical numbers to illustrate your reasoning.

3.18 Use the test scores in Exercise 1.8 two make two boxplots, one using the first-grade scores and the other the eighth-grade scores. Summarize the main differences between these two boxplots.

3.19 Exercise 2.26 shows the percentage returns for the ten Dow stocks with the highest dividend/price (*D*/*P*) ratio and the ten Dow stocks with the lowest *D*/*P* ratio. Display these data in two side-by-side boxplots. Do the data appear to have similar or dissimilar medians and dispersion?

3.20 Table 1.4 shows the U.S. dollar prices of Big Mac hamburgers in 20 countries. Use a boxplot to summarize these data.

3.21 Figure 3.11 shows two boxplots. Which data set has the higher median? The higher standard deviation?

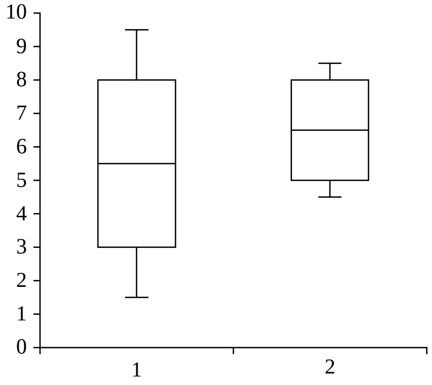

Figure 3.11
Exercise 3.21.

3.22 Figure 3.12 shows two boxplots. Which data set has the higher median? The higher mean? The higher standard deviation?

3.23 Draw a freehand sketch of two side-by-side boxplots, one the boxplot shown in Figure 3.13 and the other the same boxplot if 2 is added to the value of each observation in that boxplot.

3.24 Draw a freehand sketch of two side-by-side boxplots, one the boxplot shown in Figure 3.14 and the other the same boxplot if all the observations in that boxplot are multiplied by 2.

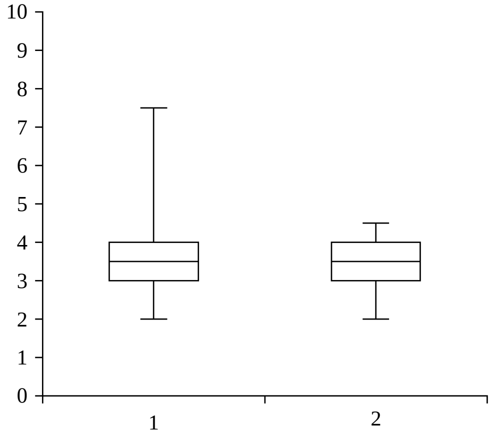

Figure 3.12
Exercise 3.22.

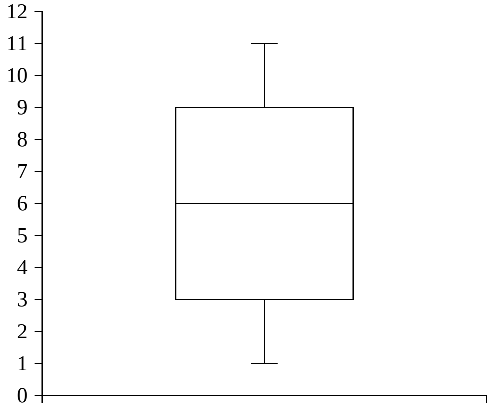

Figure 3.13
Exercise 3.23.

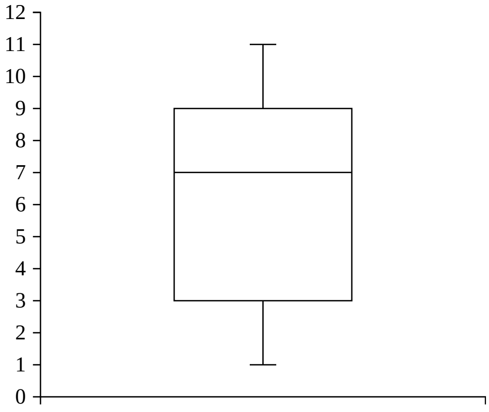

Figure 3.14
Exercise 3.24.

3.25 Here are data on the number of U.S. households (in millions) of different sizes in 2000:

Household size	1	2	3	4	5	6	7 or more
Number of households	31	39	19	16	7	3	1

a. Looking at the 116 million households, ordered by size, what is the size of the median household?
b. Looking at all 291 million individuals, ordered by the size of the household in which they live, what is the size of the household that the median individual lives in?
c. Pomona College reports its average class size as 14 students. However, a survey that asked Pomona College students the size of the classes they were enrolled in found that the average class size was 38 students and that 70 percent of the respondents' classes had more than 14 students. Use the insights gained in the first two parts of this exercise to explain this disparity.

3.26 Joanna is staring work with an annual salary of $40,000. If her salary increases by 5 percent a year, what will her salary be 40 years from now? If prices increase by 3 percent a year, how much higher will prices be 40 years from now than they are today? What will be the real value of the salary Joanna receives 40 years from now in terms of current dollars?

3.27 Treasury zeros pay a fixed amount at maturity. For example, a Treasury zero could be purchased on January 2, 2009, for $445.19 and might be worth $1,000 when it matures 25 years later. What is the annual rate of return on this zero?

3.28 A stock's price/earnings (*P*/*E*) ratio is the per-share price of its stock divided by the company's annual earnings per share. The *P*/*E* ratio for the stock market as a whole is used by some analysts as a measure of whether stocks are cheap or expensive, in comparison with other historical periods. Table 3.3 shows some annual *P*/*E* ratios for the New York Stock Exchange (NYSE). Calculate the mean and standard deviation. The stock market reached a peak in August 1987, when the Dow Jones Industrial Average topped 2,700. The Dow slipped back to 2,500 in October of 1987 and then to 2,250. Then, on a single day, October 19, 1987, the Dow fell by 508 points. At its August 1987 peak, the market's price/earnings ratio was 23. Was this *P*/*E* value more than two standard deviations above the mean *P*/*E* for 1970–1986? Was it more than 1.5 times the interquartile range above the median? Draw a box-and-whiskers diagram using these 1970–1986 *P*/*E* data.

3.29 In the early 1970s, many investors were infatuated by the Nifty 50—a group of 50 growth stocks that were thought to be so appealing that they should be bought and never sold, regardless of price. Table 3.4 shows the price/earnings (*P*/*E*) ratio in December 1972 and the annualized stock return (*R*) from December 31, 1972, through December 31, 2001, for each of these 50 stocks. Calculate the mean and median values of the returns. Explain why these are not equal.

3.30 Use the data in the preceding exercise to make side-by-side boxplots of the returns for the 25 stocks with the highest *P*/*E*s and the 25 stocks with the lowest *P*/*E*s. What differences do you observe in these boxplots?

Table 3.3: Exercise 3.28

Year	*P*/*E*	Year	*P*/*E*	Year	*P*/*E*
1970	15.5	1976	11.2	1982	8.6
1971	18.5	1977	9.3	1983	12.5
1972	18.2	1978	8.3	1984	10.0
1973	14.0	1979	7.4	1985	12.3
1974	8.6	1980	7.9	1986	16.4
1975	10.9	1981	8.4		

Table 3.4: Exercise 3.29

P/*E*	*R*	*P*/*E*	*R*	*P*/*E*	*R*	*P*/*E*	*R*
90.74	−14.68	50.47	2.45	40.77	9.78	29.34	15.55
85.67	10.50	50.37	13.19	39.00	10.30	28.97	16.99
83.26	−6.84	50.00	12.36	38.85	13.13	27.60	15.35
81.64	8.97	49.45	10.37	38.66	6.68	26.12	15.57
78.52	10.10	48.82	−1.64	38.32	3.19	26.07	12.40
75.75	5.66	48.77	0.89	37.36	9.68	25.93	17.68
65.43	6.04	48.17	1.72	36.89	7.62	25.86	14.12
62.11	−1.37	47.60	13.15	34.10	4.83	25.59	4.91
61.85	13.35	46.28	11.27	33.87	14.21	25.50	10.80
59.97	0.93	46.03	13.14	32.04	11.94	22.37	13.36
54.31	−1.07	45.94	14.27	31.90	13.55	16.28	9.99
53.12	−1.47	41.11	9.95	30.77	6.94		
51.82	11.17	41.00	10.96	30.05	14.66		

3.31 Use the data in Exercise 3.29 to make a scatter diagram with the *P*/*E* ratio on the horizontal axis and the return on the vertical axis. Does there appear to be a positive relationship, negative relationship, or essentially no relationship?

3.32 Table 3.5 shows ringer percentages for Mary Ann Peninger at the 2000 Women's world horseshoe championships, when she threw first and when she threw second. For example, in the second game, she threw 68.8 percent ringers when she went first and 50.0 percent ringers when she went second. Draw two side-by-side boxplots for her ringer percentages when she threw first and second. What differences do you notice?

3.33 Use the data in Table 3.2 to calculate the annual rate of increase in the CPI over the 10-year period from December 2000 to December 2010.

3.34 Explain the error in the conclusion reached by a security analyst [16]:

> *The Dow Jones Industrial Average peaked at 381.17 during September 3, 1929. The so-called Great Crash pushed this index down 48% by November 13. But, by April 17, 1930, the index rebounded 48% from the November bottom. In other words, anyone who bought a diversified portfolio of stocks during September 1929 would have experienced no net change in the value of his portfolio by April of 1930.*

3.35 Vincent Van Gogh sold only one painting during his lifetime, for about $30. In 1987, a sunflower still life he painted in 1888 sold for $39.85 million, more than three times the highest price paid previously for any work of art. Observers attributed the record price in part to the fact that his other sunflower paintings are all in museums and most likely will never be available for sale. If this painting had been purchased for $30 in 1888 and sold in 1987 for $39.85 million, what would have been the annual rate of return?

Table 3.5: Exercise 3.32 [15]

Game	Threw First	Threw Second
1	86.4	57.1
2	68.8	50.0
3	78.6	68.8
4	61.5	100.0
5	90.0	77.8
6	72.2	79.4
7	75.0	80.8
8	86.7	68.2
9	63.3	80.0
10	83.3	58.9
11	76.7	90.0
12	72.7	66.7
13	80.0	85.0
14	60.7	85.0
15	77.8	77.8

3.36 The world's population was 4.43 billion in 1980 and 6.85 billion in 2010. What was the annual rate of increase during these 30 years? At this rate, what will the world's population be in the year 2050?

3.37 The Dow Jones Industrial Average was 240.01 on October 1, 1928 (when the Dow expanded to 30 stocks), and 14,087.55 on October 1, 2007. The CPI was 51.3 in 1928 and 617.7 in 2007. What was the annual percentage increase in the real value of the Dow over this 79-year period?

3.38 In the 52 years between 1956 and 2008, Warren Buffett's net worth grew from \$100,000 to \$62 billion. What was his annual rate of return?

3.39 United States real (in 2007 dollars) per capita GDP was \$1,504 in 1807 and \$45,707 in 2007. What was the annual rate of growth of real per capita GDP over this 200-year period?

3.40 The CPI was 51 in 1800 and 25 in 1900. What was the annual percentage rate of change during this 100-year period?

3.41 Use the data in Table 3.2 to determine the percentage annual rate of increase in the CPI between 1990 and 2010.

3.42 A 1989 radio commercial claimed "If you have a LifeAlert security system, your chances of becoming a victim of a serious crime are 3,000 to 4,000 percent less than your neighbor's." Is this possible?

3.43 Twelve college women and men were asked to read a short article and, when they were finished, estimate how long it took them to read the article. Table 3.6 shows the actual and estimated times (both in seconds) and the percentage difference between the two. Summarize the percentage error data with two side-by-side boxplots, one for women and one for men.

3.44 Explain why this statement is incorrect: "When two assets have a +1.00 correlation, they move up and down at the same time, with the same magnitude."

3.45 Calculate the correlation between the official U.S. poverty thresholds for families of four people and the CPI in Table 3.7.

3.46 After its staff ran more than 5,000 miles in 30 different running shoes, *Consumer Reports* rated the shoes for overall performance on a scale of 0 to 100. Calculate the correlation between price and performance for female shoes and also for male shoes shown in Table 3.8. Briefly summarize your results.

3.47 The data in Table 3.9 were used in the 1964 *Surgeon General's Report* warning of the health risks associated with smoking, where the first variable is annual per capita cigarette consumption and the second variable is deaths from coronary heart disease

Table 3.6: Exercise 3.43

Women			Men		
Actual	Estimated	Error (%)	Actual	Estimated	Error (%)
173	150	−13.3	99	155	56.6
125	165	32.0	143	205	43.4
240	360	50.0	94	120	27.7
115	150	30.4	111	420	278.4
150	180	20.0	119	634	432.8
146	330	126.0	143	96	−32.9
124	190	53.2	77	185	140.3
83	104	25.3	111	240	116.2
76	250	228.9	78	125	60.3
121	240	98.3	115	200	73.9
75	58	−22.7	73	120	64.4
83	120	44.6	133	285	114.3

Table 3.7: Exercise 3.45

	Poverty Threshold	CPI
1960	$3,022	88.7
1970	$3,968	116.3
1980	$8,414	246.8
1990	$13,359	391.4
2000	$17,604	515.8
2010	$22,050	653.2

Table 3.8: Exercise 3.46 [17]

Female Shoes		Male Shoes	
Price	Rating	Price	Rating
124	82	83	85
70	80	70	85
72	78	70	84
70	77	75	82
65	75	70	78
90	73	82	72
60	71	70	72
80	70	125	70
55	70	82	68
70	69	45	68
67	67	70	65
45	64	120	64
55	62	110	64
75	57	133	62
85	46	100	50

Table 3.9: Exercise 3.47 [18]

	Cigarette Use	Heart Disease		Cigarette Use	Heart Disease
Australia	3,220	238.1	Mexico	1,680	31.9
Austria	1,770	182.1	Netherlands	1,810	124.7
Belgium	1,700	118.1	New Zealand	3,220	211.8
Canada	3,350	211.6	Norway	1,090	136.3
Denmark	1,500	144.9	Spain	1,200	43.9
Finland	2,160	233.1	Sweden	1,270	126.9
France	1,410	144.9	Switzerland	2,780	124.5
Greece	1,800	41.2	United Kingdom	2,790	194.1
Iceland	2,290	110.5	United States	3,900	259.9
Ireland	2,770	187.3	West Germany	1,890	150.3
Italy	1,510	114.3			

Table 3.10: Exercise 3.48

Year	Price Index	Change from Previous Year	Change from 2000
2005	112.2	3.4%	112%
2006	115.9	3.2%	116%
2007	119.2	2.8%	119%
2008	123.7	3.8%	134%

per 100,000 persons aged 35 to 64. Calculate the correlation. Does the value of the correlation suggest that there is a positive relationship, negative relationship, or essentially no relationship between cigarette consumption and heart disease?

3.48 Identify the error in Table 3.10, showing consumer prices and the change in prices based on a price index equal to 100 in 2000.

3.49 If $Y = 10 - 3X$, then the correlation between X and Y is

a. 1.

b. 0.

c. −1.

d. Undefined.

3.50 Explain how the following two newspaper headlines that appeared on the same day could both be accurate: "Orders for Machine Tools Increased 45.5% in October" [19]; "October Orders for Machine Tools Decreased 29%" [20].