The mean, median, and mode are complementary measures. They describe different aspects of the data. In any particular example, some or all their values may be useful. Be on the lookout for misleading statistical analyses, such as using one statistic when another would be more informative. People who present statistical conclusions often choose the statistic giving the impression they wish to convey. Recall Example 3.5 (p. 40) on Leonardo's Pizza employees, with the extreme outlying income observation. Be wary of the mean when the distribution may be highly skewed.

3.3 DESCRIBING VARIABILITY OF THE DATA

A measure of center alone is not adequate for numerically describing data for a quantitative variable. It describes a typical value, but not the spread of the data about that typical value. The two distributions in Figure 3.12 illustrate. The citizens of nation A and the citizens of nation B have the same mean annual income ($25,000). The distributions of those incomes differ fundamentally, however, nation B being much less variable. An income of $30,000 is extremely large for nation B, but not especially large for nation A. This section introduces statistics that describe the variability of a data set.

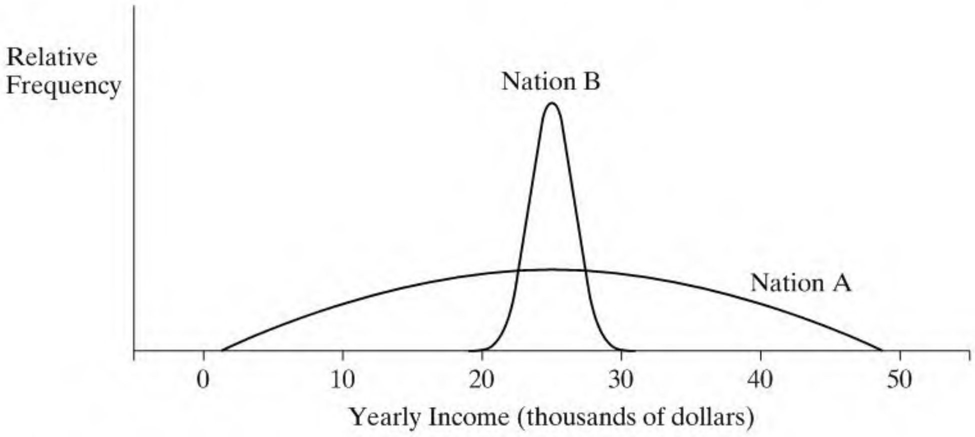

FIGURE 3.12: Distributions with the Same Mean but Different Variability

The Range

The difference between the largest and smallest observations is the simplest way to describe variability.

Range

The ***range*** is the difference between the largest and smallest observations.

For nation A, from Figure 3.12, the range of income values is about $50,000 − 0 = $50,000. For nation B, the range is about $30,000 − $20,000 = $10,000. Nation A has greater variability of incomes.

The range is not, however, sensitive to other characteristics of data variability. The three distributions in Figure 3.13 all have the same mean ($25,000) and range ($50,000), but they differ in variability about the center. In terms of distances of observations from the mean, nation A has the most variability, and nation B the least. The incomes in nation A tend to be farthest from the mean, and the incomes in nation B tend to be closest.

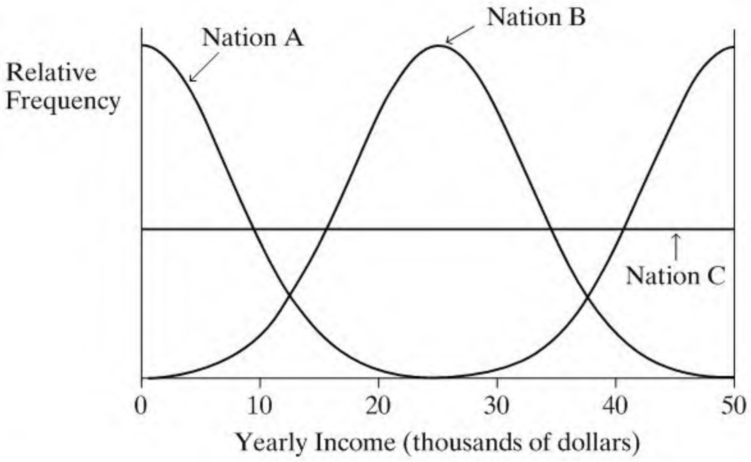

FIGURE 3.13: Distributions with the Same Mean and Range, but Different Variability about the Mean

Standard Deviation

Other measures of variability are based on the deviations of the data from a measure of center such as their mean.

Deviation

The ***deviation*** of an observation y_i from the sample mean y is $(y_i - y)$, the difference between them.

Each observation has a deviation. The deviation is *positive* when the observation falls *above* the mean. The deviation is *negative* when the observation falls *below* the mean. The interpretation of $\bar{y}$ as the center of gravity of the data implies that the sum of the positive deviations equals the negative of the sum of negative deviations. Thus, the sum of all the deviations about the mean, $\Sigma(y_i - \bar{y})$, equals 0. Because of this, measures of variability use either the absolute values or the squares of the deviations. The most popular measure uses the squares.

Standard Deviation

The ***standard deviation*** s of n observations is

$$s = \sqrt{\frac{\Sigma(y_i - \bar{y})^2}{n - 1}} = \sqrt{\frac{\text{sum of squared deviations}}{\text{sample size} - 1}}.$$

This is the positive square root of the ***variance*** s^2, which is

$$s^2 = \frac{\Sigma(y_i - \bar{y})^2}{n - 1} = \frac{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2}{n - 1}.$$

The *variance* is approximately an average of the squared deviations. The units of measurement are the squares of those for the original data, since it uses squared deviations. This makes the variance difficult to interpret. It is why we use instead its square root, the *standard deviation*.

The expression $\Sigma(y_i - \bar{y})^2$ in these formulas is called a ***sum of squares***. It represents squaring each deviation and then adding those squares. It is incorrect to first add the deviations and then square that sum; this gives a value of 0. The larger the deviations, the larger the sum of squares and the larger s tends to be.

Although its formula looks complicated, the most basic interpretation of the standard deviation s is quite simple: s is a sort of *typical distance* of an observation from the mean. So the larger the standard deviation s, the greater the spread of the data.

EXAMPLE 3.7 Comparing Variability of Quiz Scores

Each of the following sets of quiz scores for two small samples of students has a mean of 5 and a range of 10:

Sample 1:	0, 4, 4, 5, 7, 10
Sample 2:	0, 0, 1, 9, 10, 10.

By inspection, sample 1 shows less variability about the mean than sample 2. Most scores in sample 1 are near the mean of 5, whereas all the scores in sample 2 are quite far from 5.

For sample 1,

$$\Sigma (y_i - \bar{y})^2 = (0 - 5)^2 + (4 - 5)^2 + (4 - 5)^2 + (5 - 5)^2 + (7 - 5)^2 + (10 - 5)^2 = 56,$$

so the variance equals

$$s^2 = \frac{\Sigma (y_i - \bar{y})^2}{n - 1} = \frac{56}{6 - 1} = \frac{56}{5} = 11.2.$$

The standard deviation for sample 1 equals $s = \sqrt{11.2} = 3.3$. For sample 2, you can verify that $s^2 = 26.4$ and $s = \sqrt{26.4} = 5.1$. Since $3.3 < 5.1$, the standard deviations tell us that sample 1 is less variable than sample 2. ■

Statistical software and many hand calculators can find the standard deviation. You should do the calculation yourself for a few small data sets to get a feel for what this measure represents. The answer you get may differ slightly from the value reported by software, depending on how much you round off in performing the calculation.

Properties of the Standard Deviation

- $s \geq 0$.
- $s = 0$ only when all observations have the same value. For instance, if the ages for a sample of five students are 19, 19, 19, 19, 19, then the sample mean equals 19, each of the five deviations equals 0, and $s = 0$. This is the minimum possible variability.
- The greater the variability about the mean, the larger is the value of s. For example, Figure 3.5 shows that murder rates are much more variable among U.S. states than among Canadian provinces. In fact, the standard deviations are $s = 4.0$ for the United States and $s = 0.8$ for Canada.
- The reason for using $(n - 1)$, rather than n, in the denominator of s (and s^2) is a technical one regarding inference about population parameters, discussed in Chapter 5. When we have data for an entire population, we replace $(n - 1)$ by the actual population size; the population variance is then precisely the mean of the squared deviations. In that case, the standard deviation can be no larger than half the range.
- If the data are rescaled, the standard deviation is also rescaled. For instance, if we change annual incomes from dollars (such as 34,000) to thousands of dollars (such as 34.0), the standard deviation also changes by a factor of 1000 (such as from 11,800 to 11.8).

Interpreting the Magnitude of *s*

A distribution with $s = 5.1$ has greater variability than one with $s = 3.3$, but how do we interpret *how large* $s = 5.1$ is? We've seen that a rough answer is that s is a typical distance of an observation from the mean. To illustrate, suppose the first exam in your course, graded on a scale of 0 to 100, has a sample mean of 77. A value of $s = 0$ in unlikely, since every student must then score 77. A value such as $s = 50$ seems implausibly large for a typical distance from the mean. Values of s such as 8 or 12 seem much more realistic.

More precise ways to interpret s require further knowledge of the shape of the frequency distribution. The following rule provides an interpretation for many data sets.

Empirical Rule

If the histogram of the data is approximately bell shaped, then

1. About 68% of the observations fall between $\bar{y} - s$ and $\bar{y} + s$.
2. About 95% of the observations fall between $\bar{y} - 2s$ and $\bar{y} + 2s$.
3. All or nearly all observations fall between $\bar{y} - 3s$ and $\bar{y} + 3s$.

The rule is called the Empirical Rule because many distributions seen in practice (that is, *empirically*) are approximately bell shaped. Figure 3.14 is a graphical portrayal of the rule.

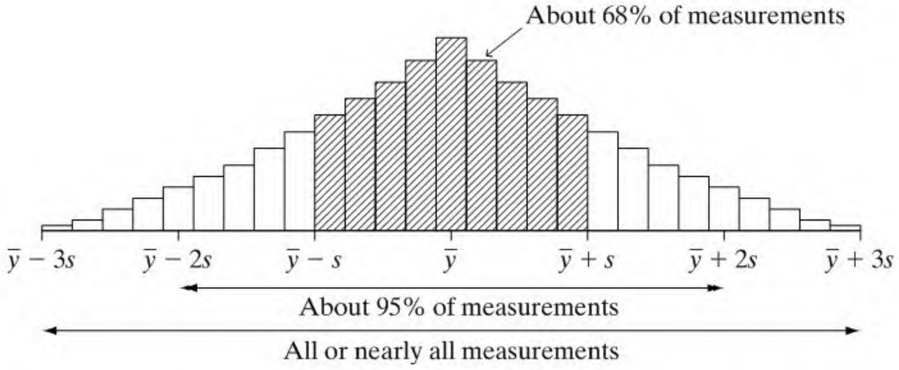

FIGURE 3.14: Empirical Rule: Interpretation of the Standard Deviation for a Bell-Shaped Distribution

EXAMPLE 3.8 Describing a Distribution of SAT Scores

The Scholastic Aptitude Test (SAT, see www.collegeboard.com) has three portions: Critical Reading, Mathematics, and Writing. For each portion, the distribution of scores is approximately bell shaped. Each portion has mean about 500 and standard deviation about 100. Figure 3.15 portrays this. By the Empirical Rule, for each portion, about 68% of the scores fall between 400 and 600, because 400 and 600 are the numbers that are *one* standard deviation below and above the mean of 500. About 95% of the scores fall between 300 and 700, the numbers that are *two* standard deviations from the mean. The remaining 5% fall either below 300 or above 700. The distribution is roughly symmetric about 500, so about 2.5% of the scores fall above 700 and about 2.5% fall below 300. ■

The Empirical Rule applies only to distributions that are approximately bell-shaped. For other shapes, the percentage falling within two standard deviations of the mean need not be near 95%. It could be as low as 75% or as high as 100%. The

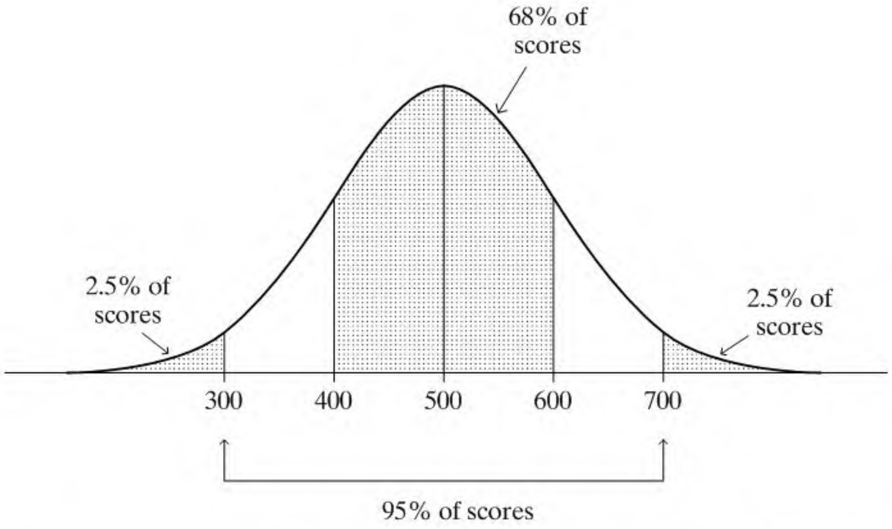

FIGURE 3.15: A Bell-Shaped Distribution of Scores for a Portion of the SAT, with Mean 500 and Standard Deviation 100

Empirical Rule may not work well if the distribution is highly skewed or if it is highly discrete, with the variable taking few values. The exact percentages depend on the form of the distribution, as the next example demonstrates.

EXAMPLE 3.9 Familiarity with AIDS Victims

A GSS asked, "How many people have you known personally, either living or dead, who came down with AIDS?" Table 3.7 shows part of a computer printout for summarizing the 1598 responses on this variable. It indicates that 76% of the responses were 0.

TABLE 3.7: Frequency Distribution of the Number of People Known Personally with AIDS

```
          AIDS   Frequency   Percent
          ---------------------------
            0       1214       76.0
            1        204       12.8
            2         85        5.3
            3         49        3.1
            4         19        1.2
            5         13        0.8
            6          5        0.3
            7          8        0.5
            8          1        0.1

N          1598
Mean       0.47
Std Dev    1.09
```

The mean and standard deviation are $\bar{y} = 0.47$ and $s = 1.09$. The values 0 and 1 both fall within one standard deviation of the mean. Now 88.8% of the distribution falls at these two points, or within $\bar{y} \pm s$. This is considerably larger than the 68% that the Empirical Rule states. The Empirical Rule does not apply to this distribution, because it is not even approximately bell shaped. Instead, it is highly skewed to the right, as you can check by sketching a histogram for Table 3.7. The smallest value in the distribution (0) is less than one standard deviation below the mean; the largest value in the distribution (8) is nearly seven standard deviations above the mean. ■

Whenever the smallest or largest observation is less than a standard deviation from the mean, this is evidence of severe skew. For instance, a recent statistics exam having scale from 0 to 100 had $\bar{y} = 86$ and $s = 15$. The upper bound of 100 was less than one standard deviation above the mean. The distribution was highly skewed to the left.

The standard deviation, like the mean, can be greatly affected by an outlier, especially for small data sets. For instance, the murder rates shown in Figure 3.5 for the 50 U.S. states have $\bar{y} = 7.3$ and $s = 4.0$. The distribution is somewhat irregular, but 68% of the states have murder rates within one standard deviation of the mean and 98% within two standard deviations. Now suppose we include the murder rate for the District of Columbia, which equaled 78.5, in the data set. Then $\bar{y} = 8.7$ and $s = 10.7$. The standard deviation more than doubles. Now 96.1% of the murder rates (all except D.C. and Louisiana) fall within one standard deviation of the mean.

3.4 MEASURES OF POSITION

Another way to describe a distribution is with a measure of ***position***. This tells us the point at which a given percentage of the data fall below (or above) that point. As special cases, some measures of position describe center and some describe variability.

Quartiles and Other Percentiles

The range uses two measures of position, the maximum value and the minimum value. The median is a measure of position, with half the data falling below it and half above it. The median is a special case of a set of measures of position called *percentiles*.

Percentile

The ***pth percentile*** is the point such that p% of the observations fall below or at that point and $(100 - p)$% fall above it.

Substituting $p = 50$ in this definition gives the 50th percentile. This is the median. The median is larger than 50% of the observations and smaller than the other $(100 - 50) = 50$%. Two other commonly used percentiles are the *lower quartile* and the *upper quartile*.

Lower and Upper Quartiles

The 25th percentile is called the ***lower quartile***. The 75th percentile is called the ***upper quartile***. One quarter of the data fall below the lower quartile. One quarter fall above the upper quartile.

The quartiles result from $p = 25$ and $p = 75$ in the percentile definition. The lower quartile is the median for the observations that fall below the median, that is, for the bottom half of the data. The upper quartile is the median for the observations that fall above the median, that is, for the upper half of the data. The quartiles together with the median split the distribution into four parts, each containing one-fourth of the observations. See Figure 3.16.