a group for which the distribution is U-shaped, for example. See Figure 3.7. In the U-shaped distribution, the highest points (representing the largest frequencies) are at the lowest and highest scores, whereas in the bell-shaped distribution, the highest point is near the middle value. A U-shaped distribution indicates a polarization on the variable between two sets of subjects. A bell-shaped distribution indicates that most subjects tend to fall near a central value.

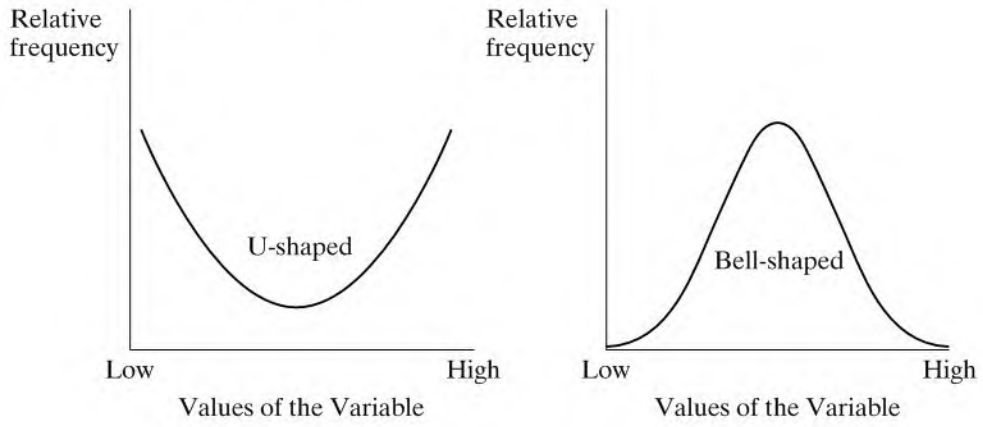

FIGURE 3.7: U-Shaped and Bell-Shaped Frequency Distributions

The distributions in Figure 3.7 are ***symmetric***: The side of the distribution below a central value is a mirror image of the side above that central value. Most distributions encountered in the social sciences are not symmetric. Figure 3.8 illustrates. The parts of the curve for the lowest values and the highest values are called the ***tails*** of the distribution. Often, as in Figure 3.8, one tail is much longer than the other. A distribution is said to be ***skewed to the right*** or ***skewed to the left***, according to which tail is longer.

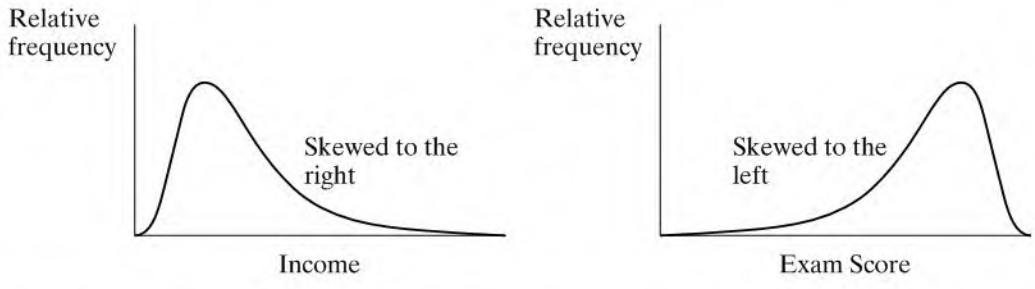

FIGURE 3.8: Skewed Frequency Distributions. The longer tail indicates the direction of skew.

To compare frequency distributions or histograms for two groups, you can give verbal descriptions using characteristics such as skew. It is also helpful to make numerical comparisons such as, "On the average, the murder rate for U.S. states is 5.4 higher than the murder rate for Canadian provinces." We now turn our attention to numerical descriptive statistics.

3.2 DESCRIBING THE CENTER OF THE DATA

This section presents statistics that describe the center of a frequency distribution for a quantitative variable. The statistics show what a *typical* observation is like.

The Mean

The best known and most commonly used measure of the center is the ***mean***.

Mean

The ***mean*** is the sum of the observations divided by the number of observations.

The mean is often called the ***average***.

EXAMPLE 3.4 Female Economic Activity in Europe

Table 3.4 shows an index of female economic activity for the countries of South America and of Eastern Europe in 2003. The number specifies female employment as a percentage of male employment. In Argentina, for instance, the number of females in the work force was 48% of the number of males in the work force. (The value was 83 in the United States and in Canada.)

TABLE 3.4: Female Economic Activity in South America and Eastern Europe; Female Employment as a Percentage of Male Employment

South America		Eastern Europe	
Country	Activity	Country	Activity
Argentina	48	Czech republic	83
Bolivia	58	Estonia	82
Brazil	52	Hungary	72
Chile	50	Latvia	80
Colombia	62	Lithuania	80
Ecuador	40	Poland	81
Guyana	51	Slovakia	84
Paraguay	44	Slovenia	81
Peru	45		
Uruguay	68		
Venezuela	55		

Source: Human Development Report 2005, United Nations Development Programme.

For the eight observations for Eastern Europe, the sum equals

$$83 + 82 + 72 + 80 + 80 + 81 + 84 + 81 = 643.$$

The mean female economic activity equals $643/8 = 80.4$. By comparison, you can check that the mean for the 11 South American countries equals $573/11 = 52.1$. Female economic activity tends to be considerably lower in South America than in Eastern Europe. ■

We use the following notation for the mean in formulas for it and for statistics that use the mean.

Notation for Observations and Sample Mean

The sample size is symbolized by n. For a variable denoted by y, its observations are denoted by $y_1, y_2, \ldots, y_n$. The sample mean is denoted by $\bar{y}$.

The symbol $\bar{y}$ for the sample mean is read as "y-bar." Throughout the text, letters near the end of the alphabet denote variables. The n sample observations on a variable y are denoted by y_1 for the first observation, y_2 the second, and so forth. For example, for female economic activity in Eastern Europe, $n = 8$ and the observations are $y_1 = 83$, $y_2 = 82, \ldots, y_8 = 81$. A bar over a letter represents the sample mean for that variable. For instance, $\bar{x}$ represents the sample mean for a variable denoted by x.

The definition of the sample mean says that

$$\bar{y} = \frac{y_1 + y_2 + \cdots + y_n}{n}.$$

The symbol Σ (uppercase Greek letter sigma) represents the process of summing. For instance, Σy_i represents the sum $y_1 + y_2 + \cdots + y_n$. This symbol stands for the sum of the y-values, where the index i represents a typical value in the range 1 to n. To illustrate, for the Eastern European data,

$$\sum y_i = y_1 + y_2 + \cdots + y_8 = 83 + 82 + \cdots + 81 = 643.$$

The symbol is sometimes even further abbreviated as Σy. Using this summation symbol, we have the shortened expression for the sample mean of n observations,

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

Properties of the Mean

Here are some properties of the mean:

- The formula for the mean uses numerical values for the observations. So the mean is appropriate only for quantitative variables. It is not sensible to compute the mean for observations on a nominal scale. For instance, for religion measured with categories such as (Protestant, Catholic, Jewish, Other), the mean religion does not make sense, even though these levels may sometimes be coded by numbers for convenience. Similarly, we cannot find the mean of observations on an ordinal rating such as excellent, good, fair, and poor, unless we assign numbers such as 4, 3, 2, 1 to the ordered levels, treating it as quantitative.
- The mean can be highly influenced by an observation that falls well above or well below the bulk of the data, called an ***outlier***.

EXAMPLE 3.5 Effect of Outlier on Mean Income

The owner of Leonardo's Pizza reports that the mean annual income of employees in the business is \$40,900. In fact, the annual incomes of the seven employees are \$11,200, \$11,400, \$11,700, \$12,200, \$12,300, \$12,500, and \$215,000. The \$215,000 income is the salary of the owner's son, who happens to be an employee. The value \$215,000 is an outlier. The mean computed for the other six observations alone equals \$11,883, quite different from the mean of \$40,900 including the outlier. ■

This example shows that the mean is not always typical of the observations in the sample. This commonly happens with small samples when at least one observation is

much larger or much smaller than the others, such as in highly skewed distributions.

- The mean is pulled in the direction of the longer tail of a skewed distribution, relative to most of the data.

  In Example 3.5, the large observation \$215,000 results in an extreme skewness to the right of the income distribution. This skewness pulls the mean above six of the seven observations. In general, the more highly skewed the distribution, the less typical the mean is of the data.

- The mean is the point of balance on the number line when an equal weight is at each observation point.

  For example, Figure 3.9 shows that if an equal weight is placed at each Eastern European observation on female economic activity from Example 3.4, then the line balances by placing a fulcrum at the point 80.4. The mean is the *center of gravity* (balance point) of the observations. This means that the sum of the distances to the mean from the observations *above* the mean equals the sum of the distances to the mean from the observations *below* the mean.

FIGURE 3.9: The Mean as the Center of Gravity, for Eastern Europe Data from Example 3.4. The line balances with a fulcrum at 80.4.

- Denote the sample means for two sets of data with sample sizes n_1 and n_2 by $\bar{y}_1$ and $\bar{y}_2$. The overall sample mean for the combined set of $(n_1 + n_2)$ observations is the ***weighted average***

$$\bar{y} = \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n_1 + n_2}.$$

  The numerator $n_1\bar{y}_1 + n_2\bar{y}_2$ is the sum of all the observations, since $n\bar{y} = \sum y$ for each set of observations. The denominator is the total sample size.

To illustrate, for the female economic activity data in Table 3.4, the South American observations have $n_1 = 11$ and $\bar{y}_1 = 52.1$, and the Eastern European observations have $n_2 = 8$ and $\bar{y}_2 = 80.4$. The overall mean economic activity for the 19 nations equals

$$\bar{y} = \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n_1 + n_2} = \frac{11(52.1) + 8(80.4)}{11 + 8} = \frac{(573 + 643)}{19} = \frac{1216}{19} = 64.$$

The weighted average of 64 is closer to 52.1, the value for South America, than to 80.4, the value for Eastern Europe. This happens because more observations come from South America than Eastern Europe.

The Median

The mean is a simple measure of the center. But other measures are also informative and sometimes more appropriate. Most important is the *median*. It splits the sample into two parts with equal numbers of observations, when they are ordered from lowest to highest.

Median

The ***median*** is the observation that falls in the middle of the ordered sample. When the sample size n is odd, a single observation occurs in the middle. When the sample size is even, two middle observations occur, and the median is the midpoint between the two.

To illustrate, the ordered income observations for the seven employees in Example 3.5 are

$$\$11{,}200, \$11{,}400, \$11{,}700, \$12{,}200, \$12{,}300, \$12{,}500, \$215{,}000.$$

The median is the middle observation, \$12,200. This is a more typical value for this sample than the sample mean of \$40,900. When a distribution is highly skewed, the median describes a typical value better than the mean.

In Table 3.4, the ordered economic activity values for the Eastern European nations are

$$72, 80, 80, 81, 81, 82, 83, 84.$$

Since $n = 8$ is even, the median is the midpoint between the two middle values, 81 and 81, which is $(81 + 81)/2 = 81$. This is close to the sample mean of 80.4, because this data set has no outliers.

The middle observation has index $(n + 1)/2$. That is, the median is the value of observation $(n + 1)/2$ in the ordered sample. When $n = 7$, $(n + 1)/2 = (7 + 1)/2 = 4$, so the median is the fourth smallest, or equivalently fourth largest, observation. When n is even, $(n + 1)/2$ falls halfway between two numbers, and the median is the midpoint of the observations with those indices. For example, when $n = 8, (n + 1)/2 = 4.5$, so the median is the midpoint between the 4th and 5th smallest observations.

EXAMPLE 3.6 Median for Grouped or Ordinal Data

Table 3.5 summarizes the distribution of the highest degree completed in the U.S. population of age 25 years and over, as estimated from the 2005 American Community Survey taken by the U.S. Bureau of the Census. The possible responses form an ordinal scale. The population size was $n = 189$ (in millions). The median score is the $(n + 1)/2 = (189 + 1)/2 = 95$th lowest. Now 30 responses fall in the first category, $(30 + 56) = 86$ in the first two, $(30 + 56 + 38) = 124$ in the first three, and so forth. The 87th to 124th lowest scores fall in category 3, which therefore contains the 95th lowest, which is the median. The median response is "Some college, no degree." Equivalently, from the percentages in the last column of the table, $(15.9\% + 29.6\%) = 45.5\%$ fall in the first two categories and $(15.9\% + 29.6\% + 20.1\%) = 65.6\%$ fall in the first three, so the 50% point falls in the third category. ■

TABLE 3.5: Highest Degree Completed, for a Sample of Americans

Highest Degree	Frequency (millions)	Percentage
Not a high school graduate	30	15.9%
High school only	56	29.6%
Some college, no degree	38	20.1%
Associate's degree	14	7.4%
Bachelor's degree	32	16.9%
Master's degree	13	6.9%
Doctorate or professional	6	3.2%

Properties of the Median

- The median, like the mean, is appropriate for quantitative variables. Since it requires only ordered observations to compute it, it is also valid for ordinal-scale data, as the previous example showed. It is not appropriate for nominal-scale data, since the observations cannot be ordered.
- For symmetric distributions, such as in Figure 3.7, the median and the mean are identical. To illustrate, the sample of observations 4, 5, 7, 9, 10 is symmetric about 7; 5 and 9 fall equally distant from it in opposite directions, as do 4 and 10. Thus, 7 is both the median and the mean.
- For skewed distributions, the mean lies toward the direction of skew (the longer tail) relative to the median. See Figure 3.10.

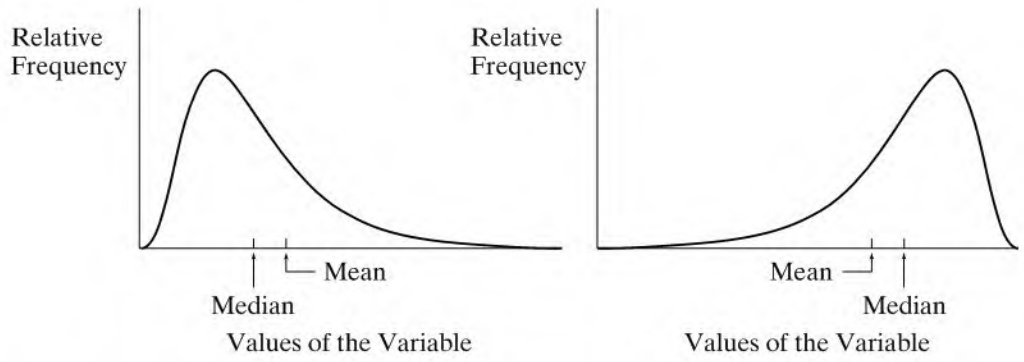

FIGURE 3.10: The Mean and the Median for Skewed Distributions. The mean is pulled in the direction of the longer tail.

For example, consider the violent crime rates of Table 3.2. The median is 36.5. The mean is $\bar{y} = 40.2$, somewhat larger than the median. Figure 3.2 showed that the violent crime rate values are skewed to the right. The mean is larger than the median for distributions that are skewed to the right. Income distributions tend to be skewed to the right. For example, household income in the United States in 2005 had a mean of about $61,000 and a median of about $44,000 (U.S. Bureau of the Census).

The distribution of grades on an exam tends to be skewed to the left when some students perform considerably poorer than the others. In this case, the mean is less than the median. For example, suppose that an exam scored on a scale of 0 to 100 has a median of 88 and a mean of 76. Then most students performed quite well (half being over 88), but apparently some scores were very much lower in order to bring the mean down to 76.

- The median is insensitive to the distances of the observations from the middle, since it uses only the ordinal characteristics of the data. For example, the following four sets of observations all have medians of 10:

Set 1:	8,	9,	10,	11,	12
Set 2:	8,	9,	10,	11,	100
Set 3:	0,	9,	10,	10,	10
Set 4:	8,	9,	10,	100,	100

- The median is not affected by outliers. For instance, the incomes of the seven employees in Example 3.5 have a median of $12,200 whether the largest observation is $20,000, $215,000, or $2,000,000.

Median Compared to Mean

The median is usually more appropriate than the mean when the distribution is highly skewed, as we observed with the Leonardo's Pizza employee incomes. The mean can be greatly affected by outliers, whereas the median is not.

For the mean we need quantitative (interval-scale) data. The median also applies for ordinal scales (see Example 3.6). To use the mean for ordinal data, we must assign scores to the categories. In Table 3.5, if we assign scores 10, 12, 13, 14, 16, 18, 20 to the categories of highest degree, representing approximate number of years of education, we get a sample mean of 13.4.

The median has its own disadvantages. For discrete data that take relatively few values, quite different patterns of data can have the same median. For instance, Table 3.6, from a GSS, summarizes the 365 female responses to the question, "How many sex partners have you had in the last 12 months?" Only six distinct responses occur, and 63.8% of those are 1. The median response is 1. To find the sample mean, to sum the 365 observations we multiply each possible value by the frequency of its occurrence, and then add. That is,

$$\sum y_i = 102(0) + 233(1) + 18(2) + 9(3) + 2(4) + 1(5) = 309.$$

The sample mean response is

$$\bar{y} = \frac{\Sigma y_i}{n} = \frac{309}{365} = 0.85.$$

If the distribution of the 365 observations among these categories were (0, 233, 18, 9, 2, 103) (i.e., we shift 102 responses from 0 to 5), then the median would still be 1, but the mean would shift to 2.2. The mean uses the numerical values of the observations, not just their ordering.

TABLE 3.6: Number of Sex Partners Last Year, for Female Respondents in GSS

Response	Frequency	Percentage
0	102	27.9
1	233	63.8
2	18	4.9
3	9	2.5
4	2	0.5
5	1	0.3

The most extreme form of this problem occurs for ***binary data***, which can take only two values, such as 0 and 1. The median equals the more common outcome, but gives no information about the relative number of observations at the two levels. For instance, consider a sample of size 5 for the variable, number of times married. The observations (1, 1, 1, 1, 1) and the observations (0, 0, 1, 1, 1) both have a median of 1. The mean is 1 for (1, 1, 1, 1, 1) and 3/5 for (0, 0, 1, 1, 1). *When observations take values of only 0 or 1, the mean equals the proportion of observations that equal 1.* Generally, for highly discrete data, the mean is more informative than the median.

In summary,

- If a distribution is highly skewed, the median is usually preferred because it better represents what is typical.

- If the distribution is close to symmetric or only mildly skewed or if it is discrete with few distinct values, the mean is usually preferred, because it uses the numerical values of all the observations.

The Mode

Another measure, the *mode*, indicates the most common outcome.

Mode

The ***mode*** is the value that occurs most frequently.

The mode is most commonly used with highly discrete variables, such as with categorical data. In Table 3.5, on the highest degree completed, for instance, the mode is "High school only," since the frequency for that category is higher than the frequency for any other rating. In Table 3.6, on the number of sex partners in the last year, the mode is 1.

Properties of the Mode

- The mode is appropriate for all types of data. For example, we might measure the mode for religion in Australia (nominal scale), for the rating given a teacher (ordinal scale), or for the number of years of education completed by Hispanic Americans (interval scale).
- A frequency distribution is called ***bimodal*** if two distinct mounds occur in the distribution. Bimodal distributions often occur with attitudinal variables when populations are polarized, with responses tending to be strongly in one direction or another. For instance, Figure 3.11 shows the relative frequency distribution of responses in a General Social Survey to the question, "Do you personally think it is wrong or not wrong for a woman to have an abortion if the family has a very low income and cannot afford any more children?" The relative frequencies in the two extreme categories are higher than those in the middle categories.

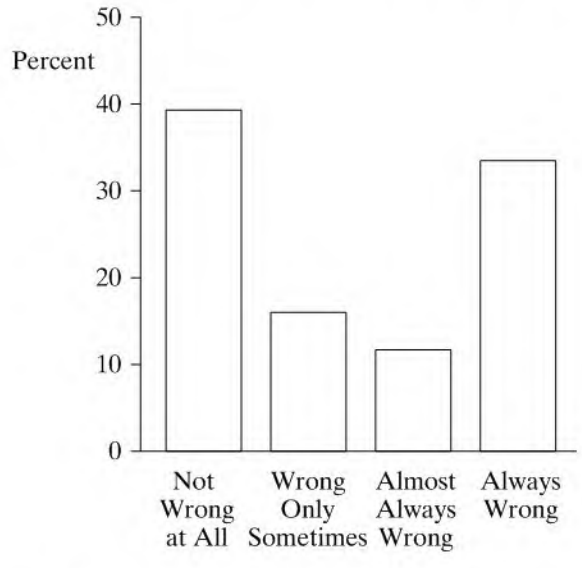

FIGURE 3.11: Bimodal Distribution for Opinion about Whether Abortion Is Wrong

- The mean, median, and mode are identical for a unimodal, symmetric distribution, such as a bell-shaped distribution.

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.

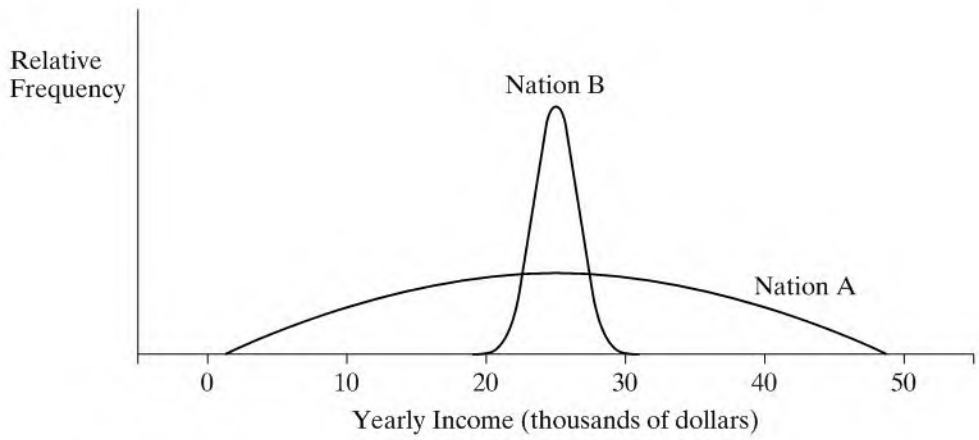

FIGURE 3.12: Distributions with the Same Mean but Different Variability

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.