CHAPTER 3

Descriptive Statistics

We've seen that statistical methods are *descriptive* or *inferential.* The purpose of descriptive statistics is to summarize data, to make it easier to assimilate the information. This chapter presents basic methods of descriptive statistics.

We first present tables and graphs that describe the data by showing the number of times various outcomes occurred. Quantitative variables also have two key features to describe numerically:

- The ***center*** of the data—a typical observation
- The ***variability*** of the data—the spread around the center

We'll learn how to describe quantitative data with statistics that summarize the center, statistics that summarize the variability, and finally with statistics that specify certain positions in the data set that summarize both center and variability.

3.1 DESCRIBING DATA WITH TABLES AND GRAPHS

Tables and graphs are useful for all types of data. We'll begin with categorical variables.

Relative Frequencies: Categorical Data

For categorical data, we list the categories and show the frequency (the number of observations) in each category. To make is easier to compare different categories, we also report proportions or percentages, also called ***relative frequencies.***

Relative Frequency

The ***relative frequency*** for a category is the ***proportion*** or ***percentage*** of the observations that fall in that category.

The *proportion* equals the number of observations in a category divided by the total number of observations. It is a number between 0 and 1 that expresses the share of the observations in that category. The *percentage* is the proportion multiplied by 100.

EXAMPLE 3.1 Household Structure in the U.S.

Table 3.1 lists the different types of households in the United States in 2005. Of 111.1 million households, for example, 24.1 million were a married couple with children. The proportion 24.1/111.1 = 0.22 were a married couple with children.

TABLE 3.1: U.S. Household Structure, 2005

Type of Family	Number (millions)	Proportion	Percentage
Married couple with children	24.1	0.22	22
Married couple, no children	31.1	0.28	28
Single householder, no spouse	19.1	0.17	17
Living alone	30.1	0.27	27
Other households	6.7	0.06	6
Total	111.1	1.00	100

Source: U.S. Census Bureau, *2005 American Community Survey*, Tables B11001, C11003.

A percentage is the proportion multiplied by 100. That is, the decimal place is moved two positions to the right. For example, since 0.22 is the proportion of families that are married couples with children, the percentage is 100(0.22) = 22%. Table 3.1 shows the proportions and percentages for all the categories. ■

The sum of the proportions equals 1.00. The sum of the percentages equals 100. (In practice, the values may sum to a slightly different number, such as 99.9 or 100.1, because of rounding.)

It is sufficient in such a table to report the percentages (or proportions) and the total sample size, since each frequency equals the corresponding proportion multiplied by the total sample size. For instance, the frequency of married couples with children equals 0.22(111.1) = 24 million. When presenting the percentages but not the frequencies, always also include the total sample size.

Frequency Distributions and Bar Graphs: Categorical Data

Table 3.1 lists the categories for household structure and the number of households of each type. Such a listing is called a *frequency distribution.*

Frequency Distribution

A ***frequency distribution*** is a listing of possible values for a variable, together with the number of observations at each value. A corresponding ***relative frequency distribution*** lists the possible values together with their proportions or percentages.

To construct a frequency distribution for a categorical variable, list the categories and count the number of observations in each.

To more easily get a feel for the data, it's helpful to look at a graph of the relative frequency distribution. A ***bar graph*** has a rectangular bar drawn over each category. The height of the bar shows the relative frequency in that category. Figure 3.1 is a bar graph for the data in Table 3.1. The bars are separated to emphasize that the variable is categorical rather than quantitative. Since household structure is a nominal variable, there is no particular natural order for the bars. The order of presentation for an ordinal variable is the natural ordering of the categories.

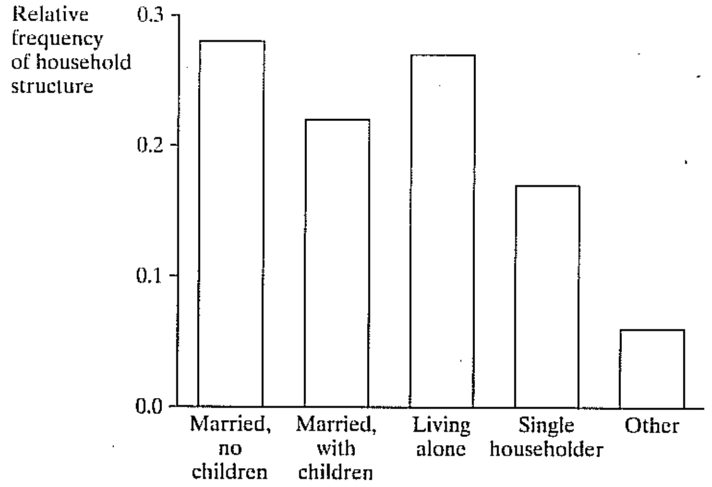

FIGURE 3.1: Relative Frequency of U.S. Household Structure Types, 2005

Another type of graph, the *pie chart*, is a circle having a "slice of the pie" for each category. The size of a slice represents the percentage of observations in the category. The bar graph is more precise than the pie chart for visual comparison of categories with similar relative frequencies.

Frequency Distributions: Quantitative Data

Frequency distributions and graphs also are useful for quantitative variables. The next example illustrates.

EXAMPLE 3.2 Statewide Violent Crime Rates

Table 3.2 lists all 50 states in the United States and their 2005 violent crime rates. This rate measures the number of violent crimes in that state in 2005 per 10,000 population. For instance, if a state had 12,000 violent crimes and a population size of 2,300,000, its violent crime rate was (12,000/2,300,000) × 10,000 = 52. It is difficult to learn much by simply reading through the violent crime rates. Tables, graphs, and numerical measures help us more fully absorb the information in these data.

First, we can summarize the data with a frequency distribution. To do this, we divide the measurement scale for violent crime rate into a set of intervals and count the number of observations in each interval. Here, we use the intervals {0–11, 12–23, 24–35, 36–47, 48–59, 60–71, 72–83}. The values Table 3.2 reports were rounded, so for example the interval 12-23 represents values between 11.5 and 23.5. Counting the number of states with violent crime rates in each interval, we get the frequency distribution shown in Table 3.3. We see that considerable variability exists in the violent crime rates.

Table 3.3 also shows the relative frequencies, using proportions and percentages. For example, 3/50 = 0.06 is the proportion for the interval 0–11, and 100(0.06) = 6 is the percentage. As with any summary method, we lose some information as the cost of achieving some clarity. The frequency distribution does not identify which states have low or high violent crime rates, nor are the exact violent crime rates known. ■

The intervals of values in frequency distributions are usually of equal width. The width equals 12 in Table 3.3. The intervals should include all possible values of the

TABLE 3.2: List of States with Violent Crime Rates Measured as Number of Violent Crimes per 10,000 Population

Alabama	43	Louisiana	65	Ohio	33
Alaska	59	Maine	11	Oklahoma	51
Arizona	51	Maryland	70	Oregon	30
Arkansas	46	Massachusetts	47	Pennsylvania	40
California	58	Michigan	51	Rhode Island	29
Colorado	34	Minnesota	26	South Carolina	79
Connecticut	31	Mississippi	33	South Dakota	17
Delaware	66	Missouri	47	Tennessee	69
Florida	73	Montana	36	Texas	55
Georgia	45	Nebraska	29	Utah	25
Hawaii	27	Nevada	61	Vermont	11
Idaho	24	New Hampshire	15	Virginia	28
Illinois	56	New Jersey	37	Washington	35
Indiana	35	New Mexico	66	West Virginia	26
Iowa	27	New York	46	Wisconsin	22
Kansas	40	North Carolina	46	Wyoming	26
Kentucky	26	North Dakota	8		

TABLE 3.3: Frequency Distribution and Relative Frequency Distribution for Violent Crime Rates

Violent Crime Rate	Frequency	Relative Frequency	Percentage
0–11	3	0.06	6
12–23	3	0.06	6
24–35	18	0.36	36
36–47	11	0.22	22
48–59	7	0.14	14
60–71	6	0.12	12
72–83	2	0.04	4
Total	50	1.00	100.0

variable. In addition, any possible value must fit into one and only one interval; that is, they should be ***mutually exclusive***.

Histograms

A graph of a relative frequency distribution for a quantitative variable is called a ***histogram***. Each interval has a bar over it, with height representing the number of observations in that interval. Figure 3.2 is a histogram for the violent crime rates.

Choosing intervals for frequency distributions and histograms is primarily a matter of common sense. If too few intervals are used, too much information is lost. For example, Figure 3.3 is a histogram of violent crime rates using the intervals 0–29, 30–59, 60–89. This is too crude to be very informative. If too many intervals are used, they are so narrow that the information presented is difficult to digest, and the histogram may be irregular and the overall pattern of the results may be obscured. Ideally, two observations in the same interval should be similar in a practical sense. To summarize annual income, for example, if a difference of $5000 in income is not considered practically important, but a difference of $15,000 is notable, we might choose intervals of width less than $15,000, such as $0–$9999, $10,000–$19,999, $20,000–$29,999, and so forth. Statistical software can automatically choose intervals for us and construct frequency distributions and histograms.

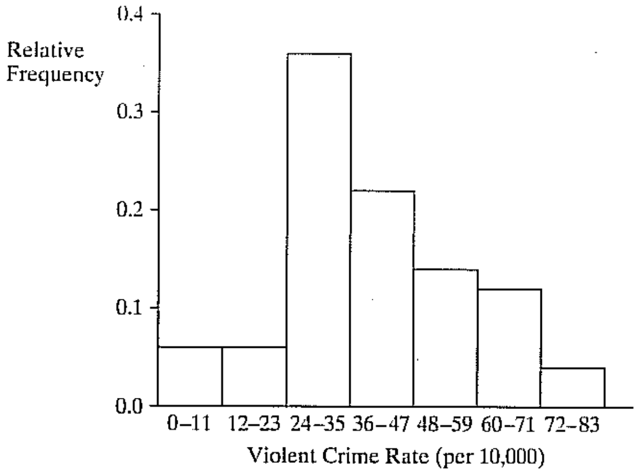

FIGURE 3.2: Histogram of Relative Frequencies for Statewide Violent Crime Rates

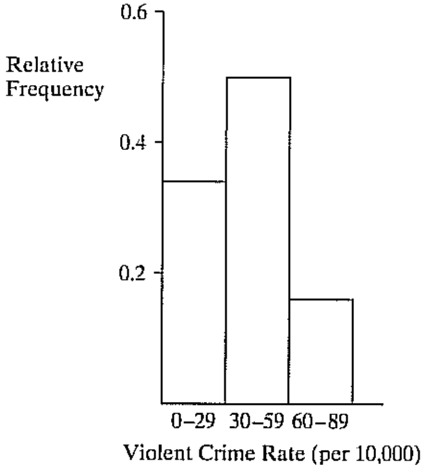

FIGURE 3.3: Histogram of Relative Frequencies for Violent Crime Rates, Using Too Few Intervals

For a discrete variable with relatively few values, a histogram has a separate bar for each possible value. For a continuous variable or a discrete variable with many possible values, you need to divide the possible values into intervals, as we did with the violent crime rates.

Stem-and-Leaf Plots

Figure 3.4 shows an alternative graphical representation of the violent crime rate data. This figure, called a ***stem-and-leaf plot***, represents each observation by its leading digit(s) (the *stem*) and by its final digit (the *leaf*). Each stem is a number to the left of the vertical bar and a leaf is a number to the right of it. For instance, on the second line, the stem of 1 and the leaves of 1, 1, 5, and 7 represent the violent crime rates 11, 11, 15, 17. The plot arranges the leaves in order on each line, from smallest to largest.

Stem	Leaf
0	8
1	1 1 5 7
2	2 4 5 6 6 6 6 7 7 8 9 9
3	0 1 3 3 4 5 5 6 7
4	0 0 3 5 6 6 6 7 7
5	1 1 1 5 6 8 9
6	1 5 6 6 9
7	0 3 9

FIGURE 3.4: Stem-and-Leaf Plot for Violent Crime Rate Data in Table 3.2

A stem-and-leaf plot conveys similar information as a histogram. Turned on its side, it has the same shape as the histogram. In fact, since the stem-and-leaf plot shows each observation, it displays information that is lost with a histogram. From Figure 3.4, the largest violent crime rate was 79 and the smallest was 8 (shown as 08 with a stem of 0 and leaf of 8). It is not possible to determine these exact values from the histogram in Figure 3.2.

Stem-and-leaf plots are useful for quick portrayals of small data sets. As the sample size increases, you can accommodate the increase in leaves by splitting the stems. For instance, you can list each stem twice, putting leaves of 0 to 4 on one line and leaves of 5 to 9 on another. When a number has several digits, it is simplest for graphical portrayal to drop the last digit or two. For instance, for a stem-and-leaf plot of annual income in thousands of dollars, a value of $27.1 thousand has a stem of 2 and a leaf of 7 and a value of $106.4 thousand has a stem of 10 and leaf of 6.

Comparing Groups

Many studies compare different groups on some variable. Relative frequency distributions, histograms, and stem-and-leaf plots are useful for making comparisons.

EXAMPLE 3.3 Comparing Canadian and U.S. Murder Rates

Stem-and-leaf plots can provide visual comparisons of two small samples on a quantitative variable. For ease of comparison, the results are plotted "back to back." Each plot uses the same stem, with leaves for one sample to its left and leaves for the other sample to its right. To illustrate, Figure 3.5 shows back-to-back stem and leaf plots of recent murder rates (measured as the number of murders per 100,000 population) for the 50 states in the U.S. and for the provinces of Canada. From this figure, it is clear that the murder rates tended to be much lower in Canada, varying between 0.7 (Prince Edward Island) and 2.9 (Manitoba) whereas those in the U.S. varied between 1.6 (Maine) and 20.3 (Louisiana). ■

Population Distribution and Sample Data Distribution

Frequency distributions and histograms apply both to a population and to samples from that population. The first type is called the ***population distribution***, and the second type is called a ***sample data distribution***. In a sense, the sample data distribution is a blurry photo of the population distribution. As the sample size increases, the sample proportion in any interval gets closer to the true population proportion. Thus, the sample data distribution looks more like the population distribution.

Canada	Stem	United States
7	0	
3 2 1	1	6 7
9 7 6 3 2 0	2	0 3 9
	3	0 1 4 4 4 6 8 9 9 9
	4	4 6
	5	0 2 3 8
	6	0 3 4 6 8 9
	7	5
	8	0 3 4 6 9
	9	0 8
	10	2 2 3 4
	11	3 3 4 4 6 9
	12	7
	13	1 3 5
	14	
	15	
	16	
	17	
	18	
	19	
	20	3

FIGURE 3.5: Back-to-Back Stem-and-Leaf Plots of Murder Rates from U.S. and Canada. Both share the same stems, with Canada leafs to the left and U.S. leafs to the right.

For a continuous variable, imagine the sample size increasing indefinitely, with the number of intervals simultaneously increasing, so their width narrows. Then, the shape of the sample histogram gradually approaches a smooth curve. This text uses such curves to represent population distributions. Figure 3.6 shows two sample histograms, one based on a sample of size 100 and the second based on a sample of size 500, and also a smooth curve representing the population distribution. Even if a variable is discrete, a smooth curve often approximates well the population distribution, especially when the number of possible values of the variable is large.

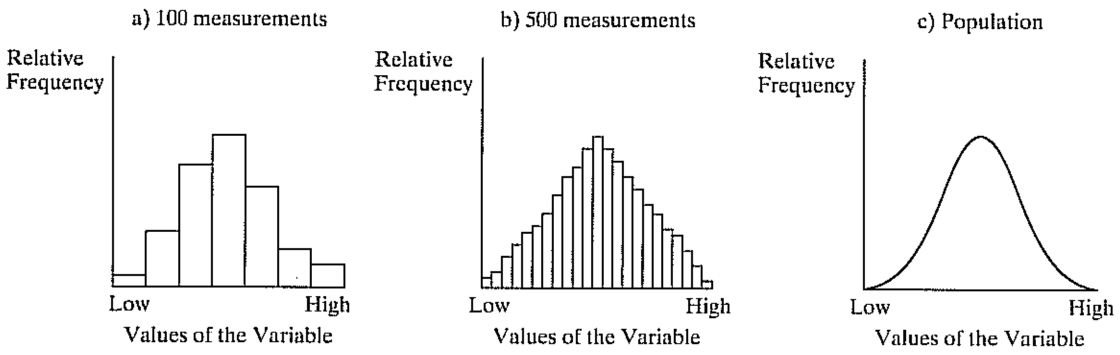

FIGURE 3.6: Histograms for a Continuous Variable. We use smooth curves to represent population distributions for continuous variables.

The Shape of a Distribution

One way to summarize a sample or a population distribution is to describe its shape. A group for which the distribution is bell-shaped is fundamentally different from 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.

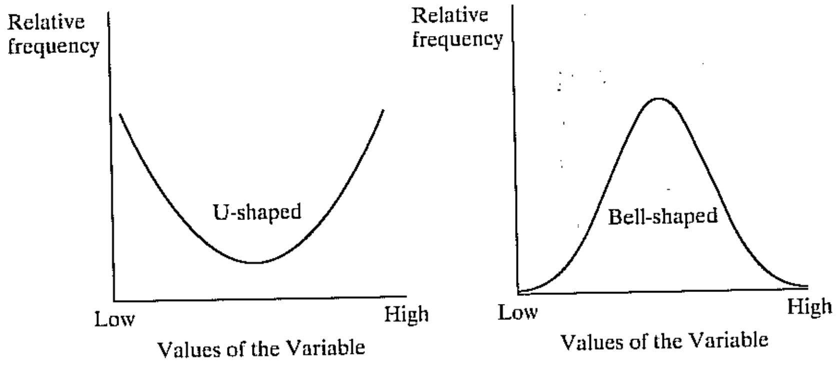

FIGURE 3.7: U-Shaped and Bell-Shaped Frequency Distributions

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.

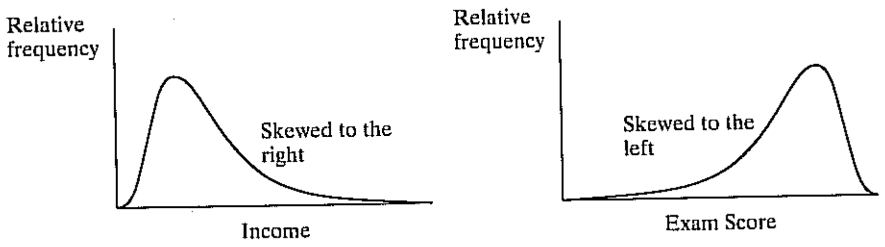

FIGURE 3.8: Skewed Frequency Distributions. The longer tail indicates the direction of skew.

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.

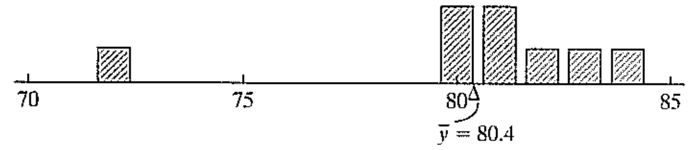

FIGURE 3.9: The Mean as the Center of Gravity, for Eastern Europe Data from Example 3.4. The line balances with a fulcrum at 80.4.

- 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 $n\bar{y} = \Sigma y$ 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.

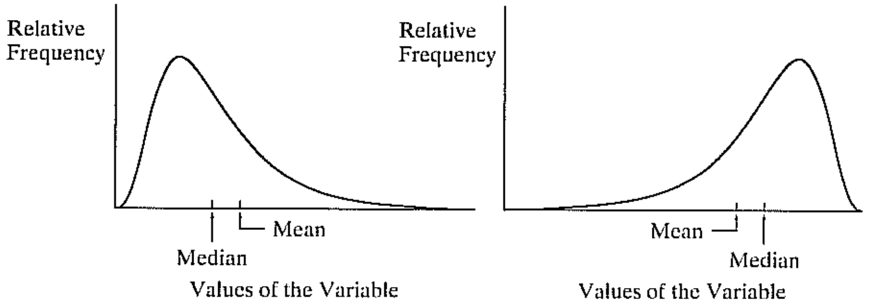

FIGURE 3.10: The Mean and the Median for Skewed Distributions. The mean is pulled in the direction of the longer tail.

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.

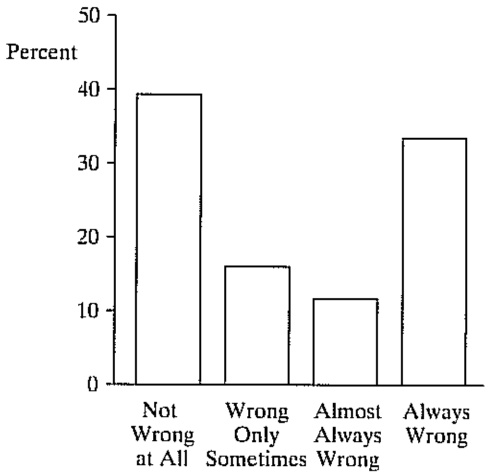

FIGURE 3.11: Bimodal Distribution for Opinion about Whether Abortion Is Wrong

- 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.

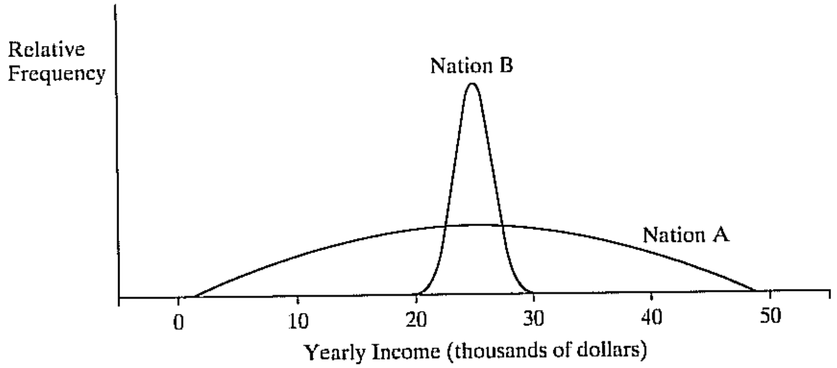

FIGURE 3.12: Distributions with the Same Mean but Different Variability

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.

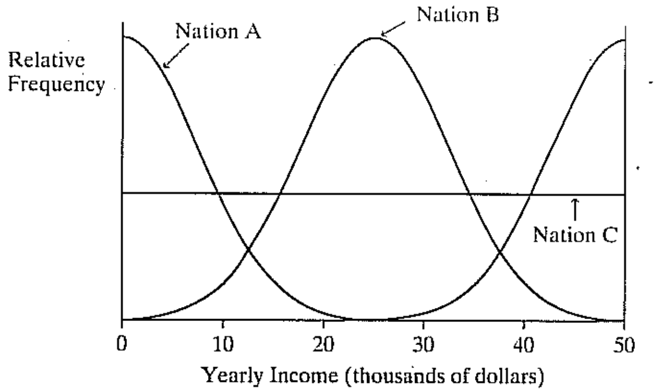

FIGURE 3.13: Distributions with the Same Mean and Range, but Different Variability about the Mean

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 $\bar{y}$ is $(y_i - \bar{y})$, 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:

$$\text{Sample 1: } 0, 4, 4, 5, 7, 10$$
$$\text{Sample 2: } 0, 0, 1, 9, 10, 10.$$

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.

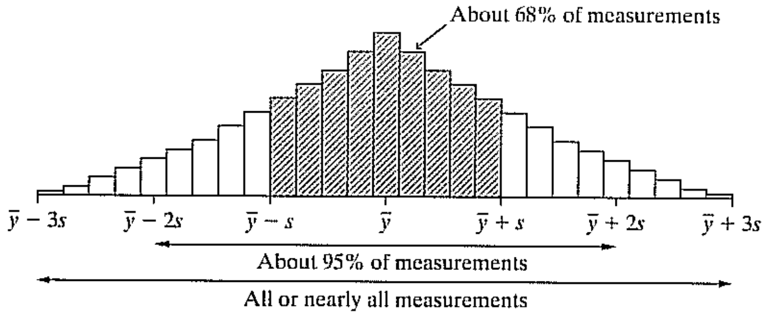

FIGURE 3.14: Empirical Rule: Interpretation of the Standard Deviation for a Bell-Shaped Distribution

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

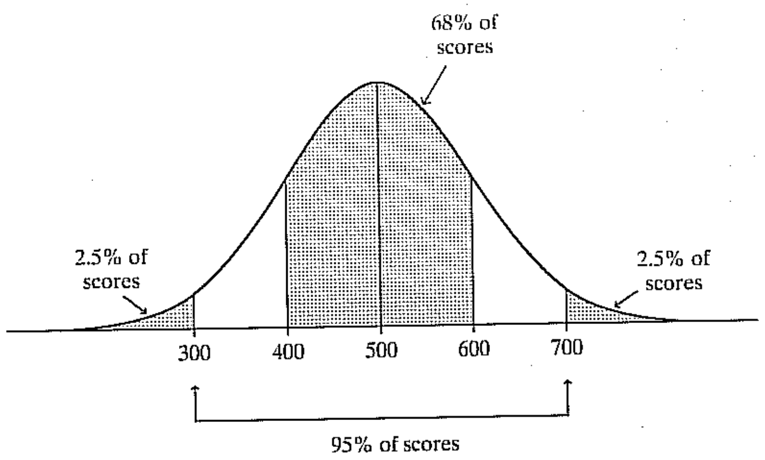

FIGURE 3.15: A Bell-Shaped Distribution of Scores for a Portion of the SAT, with Mean 500 and Standard Deviation 100

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.

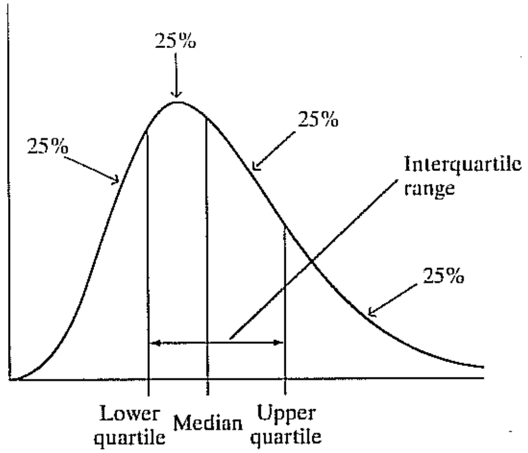

FIGURE 3.16: The Quartiles and the Interquartile Range

For the violent crime rates in Table 3.2, the sample size is $n = 50$ and the median equals 36.5. As with the median, the quartiles can easily be found from the stem-and-leaf plot of the data (Figure 3.4), which was

Stem	Leaf
0	8
1	1 1 5 7
2	2 4 5 6 6 6 6 7 7 8 9 9
3	0 1 3 3 4 5 5 6 7
4	0 0 3 5 6 6 6 7 7
5	1 1 1 5 6 8 9
6	1 5 6 6 9
7	0 3 9

The lower quartile is the median for the 25 observations below the median. It is the 13th smallest observation, or 27. The upper quartile is the median for the 25 observations above the median. It is the 13th largest observation, or 51.

In summary, since

$$\text{lower quartile} = 27, \text{ median} = 36.5, \text{ upper quartile} = 51,$$

roughly a quarter of the states had violent crime rates (i) below 27, (ii) between 27 and 36.5, (iii) between 36.5 and 51, and (iv) above 51. The distance between the upper quartile and the median, $51 - 36.5 = 14.5$, exceeds the distance $36.5 - 27 = 9.5$ between the lower quartile and the median. This commonly happens when the distribution is skewed to the right.

Software can easily find quartiles as well as other percentiles. In practice, percentiles other than the median are usually not reported for small data sets.

Measuring Variability: Interquartile Range

The difference between the upper and lower quartiles is called the ***interquartile range***, denoted by IQR. This measure describes the spread of the middle half of the observations. For the U.S. violent crime rates in Table 3.2, the interquartile range $\text{IQR} = 51 - 27 = 24$. The middle half of the murder rates fall within a range of 24. Like the range and standard deviation, the IQR increases as the variability increases, and it is useful for comparing variability of different groups. For example, 12 years earlier in 1993, the quartiles of the U.S. statewide violent crime rates were 33 and 77, giving an IQR of $77 - 33 = 44$ and showing quite a bit more variability.

An advantage of the IQR over the ordinary range or the standard deviation is that it is not sensitive to outliers. The U.S. violent crime rates range from 8 to 79, so the range is 71. When we include the observation for D.C., which was 161, the IQR changes only from 24 to 28. By contrast, the range changes from 71 to 153.

For bell-shaped distributions, the distance from the mean to either quartile is about 2/3rd of a standard deviation. Then IQR is roughly $(4/3)s$. The insensitivity of the IQR to outliers has recently increased its popularity, although in practice the standard deviation is still much more common.

Box Plots: Graphing a Five-Number Summary of Positions

The median, the quartiles, and the maximum and minimum are five positions often used as a set to describe center and spread. For instance, software reports the following five-number summary for the violent crime rates (where Q1 = lower quartile, Q3 = upper quartile, regarding the median as the second quartile):

```
100% Max    79.0
 75% Q3     51.0
 50% Med    36.5
 25% Q1     27.0
  0% Min     8.0
```

The five-number summary provides a simple description of the data. It is the basis of a graphical display called the ***box plot*** that summarizes both the center and the variability. The *box* of a box plot contains the central 50% of the distribution, from the lower quartile to the upper quartile. The median is marked by a line drawn within the box. The lines extending from the box are called *whiskers*. These extend to the maximum and minimum, except for outliers, which are marked separately.

Figure 3.17 shows the box plot for the violent crime rates, in the format provided with SPSS software. The upper whisker and upper half of the central box are longer than the lower ones. This indicates that the right tail of the distribution, which corresponds to the relatively large values, is longer than the left tail. The plot reflects the skewness to the right of violent crime rates. (Some software also plots the mean on the box plot, representing it by a + sign.)

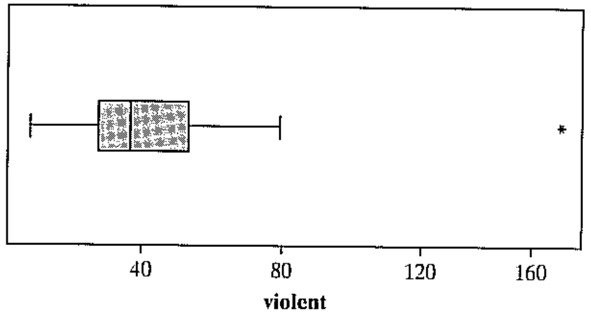

FIGURE 3.17: Box Plot of Violent Crime Rates of U.S. States and D.C.

Side-by-side box plots are useful for comparing two distributions. Figure 3.5 showed side-by-side stem-and-leaf plots of U.S. and Canadian murder rates. Figure 3.18 shows

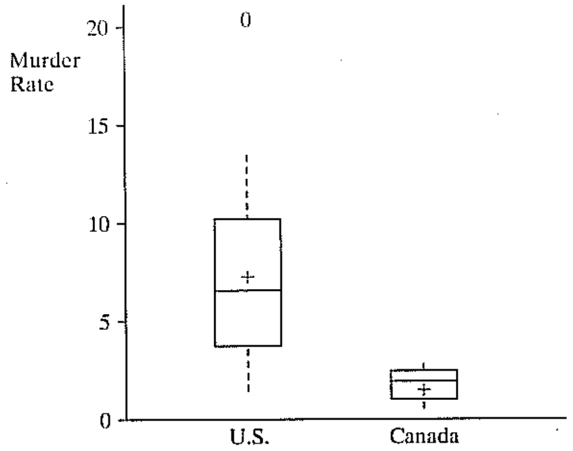

FIGURE 3.18: Box Plots for U.S. and Canadian Murder Rates

the side-by-side box plots. These side-by-side box plots reveal that the murder rates in the U.S. tend to be much higher and have much greater variability.

Outliers

Box plots identify outliers separately. To explain this, we now present a formal definition of an outlier.

Outlier

An observation is an ***outlier*** if it falls more than 1.5(IQR) above the upper quartile or more than 1.5(IQR) below the lower quartile.

In box plots, the whiskers extend to the smallest and largest observations only if those values are not outliers; that is, if they are no more than 1.5(IQR) beyond the quartiles. Otherwise, the whiskers extend to the most extreme observations within 1.5(IQR), and the outliers are marked separately. For instance, the statistical software SAS marks by an O (O for outlier) a value between 1.5 and 3.0(IQR) from the box and by an asterisk (*) a value even farther away.

Figure 3.18 shows one outlier for the U.S. with a very high murder rate. This is the murder rate of 20.3 (for Louisiana). For these data, the lower quartile = 3.9 and upper quartile = 10.3, so $IQR = 10.3 - 3.9 = 6.4$. Thus,

$$\text{Upper quartile} + 1.5(\text{IQR}) = 10.3 + 1.5(6.4) = 19.9.$$

Since $20.3 > 19.9$, the box plot highlights the observation of 20.3 as an outlier.

Why highlight outliers? It can be informative to investigate them. Was the observation perhaps incorrectly recorded? Was that subject fundamentally different from the others in some way? Often it makes sense to repeat a statistical analysis without an outlier, to make sure the conclusions are not overly sensitive to a single observation. Another reason to show outliers separately in a box plot is that they do not provide much information about the shape of the distribution, especially for large data sets.

In practice, the 1.5(IQR) criterion for an outlier is somewhat arbitrary. It is better to regard an observation satisfying this criterion as a *potential* outlier rather than a definite outlier. When a distribution has a long right tail, some observations may fall more than 1.5 IQR above the upper quartile even if they are not separated far from the bulk of the data.

How Many Standard Deviations from the Mean? The *z*-Score

Another way to measure position is by the number of standard deviations that a point falls from the mean. For example, the U.S. murder rates shown in the box plot in Figure 3.18 have a mean of 7.3 and a standard deviation of 4.0. The value of 20.3 for Louisiana falls $20.3 - 7.3 = 13.0$ above the mean. Now, 13 is $13/4 = 3.25$ standard deviations. The Louisiana murder rate is 3.25 standard deviations above the mean.

The number of standard deviations that an observation falls from the mean is called its ***z*-score**. For the murder rates of Figure 3.18, Louisiana has a z-score of

$$z = \frac{20.3 - 7.3}{4.0} = \frac{\text{Observation} - \text{Mean}}{\text{Standard Deviation}} = 3.25.$$

By the Empirical Rule, for a bell-shaped distribution it is very unusual for an observation to fall more than three standard deviations from the mean. An alternative criterion regards an observation as an outlier if it has a z-score larger than 3 in absolute value. By this criterion, the murder rate for Louisiana is an outlier.

We'll study z-scores in more detail in the next chapter. We'll see they are especially useful for bell-shaped distributions.

3.5 BIVARIATE DESCRIPTIVE STATISTICS

In this chapter we've learned how to summarize categorical and quantitative variables graphically and numerically. In the next three chapters we'll learn about basic ideas of statistical inference for a categorical or quantitative variable. Most studies have more than one variable, however, and Chapters 7–16 present methods that can handle two or more variables at a time.

Association between Response and Explanatory Variables

With multivariable analyses, the main focus is on studying *associations* among the variables. There is said to be an ***association*** between two variables if certain values of one variable tend to go with certain values of the other.

For example, consider "religious affiliation," with categories (Protestant, Catholic, Other) and "ethnic group," with categories (Anglo-American, African-American, Hispanic). In the United States, Anglo-Americans are more likely to be Protestant than are Hispanics, who are overwhelmingly Catholic. African-Americans are even more likely to be Protestant. An association exists between religious affiliation and ethnic group, because the proportion of people having a particular religious affiliation changes as ethnic group changes.

An analysis of association between two variables is called a ***bivariate*** analysis, because there are two variables. Usually one is an outcome variable on which comparisons are made at levels of the other variable. The outcome variable is called the ***response variable***. The variable that defines the groups is called the ***explanatory variable***. The analysis studies how the outcome on the response variable *depends on* or is *explained by* the value of the explanatory variable. For example, when we describe how religious affiliation depends on ethnic group, religious affiliation is the response variable. In a comparison of men and women on income, income is the response variable and gender is the explanatory variable. Income may depend on gender, not gender on income.

Often, the response variable is called the ***dependent variable*** and the explanatory variable is called the ***independent variable***. The terminology *dependent variable* refers to the goal of investigating the degree to which the response on that variable *depends on* the value of the other variable. We prefer not to use these terms, since *independent* and *dependent* are used for so many other things in statistical methods.

Comparing Two Groups Is a Bivariate Analysis

Chapter 7 will present descriptive and inferential methods for comparing two groups. For example, suppose we'd like to know whether men or women have more good friends, on the average. A GSS reports (for variable NUMFREND) that the mean number of good friends is 7.0 for men (s = 8.4) and 5.9 for women (s = 6.0). The two distributions have similar appearance, both being skewed to the right and with a median of 4.

Here, this is an analysis of two variables—number of good friends and gender. The response variable, number of good friends, is quantitative. The explanatory variable, gender, is categorical. In this case, it's common to compare means on the response variable for the categories of the categorical variable. Graphs are also useful, such as side-by-side box plots.

Bivariate Categorical Data

Chapter 8 will present methods for analyzing association between two categorical variables. Table 3.8 is an example of such data. This table results from answers to two questions on the 2006 General Social Survey. One asked whether homosexual relations are wrong. The other asked about the fundamentalism/liberalism of the respondent's religion. A table of this kind, called a ***contingency table***, displays the number of subjects observed at combinations of possible outcomes for the two variables. It displays how outcomes of a response variable are *contingent* on the category of the explanatory variable.

TABLE 3.8: Cross-Classification of Religion and Opinion about Homosexual Relations

	Opinion about Homosexual Relations				
Religion	Always Wrong	Almost Always Wrong	Sometimes Wrong	Not Wrong at All	Total
Fundamentalist	416	26	22	83	547
Liberal	213	29	52	292	586

Table 3.8 has eight possible combinations of responses. (Another possible outcome, "moderate" for the religion variable, is not shown here.) We could list the categories in a frequency distribution or construct a bar graph. Usually, though, it's more informative to do this for the categories of the response variable, separately for each category of the explanatory variable. For example, if we treat opinion about homosexual relations as the response variable, we could report the percentages in the four categories for homosexual relations, separately for each religious category.

Consider those who report being fundamentalist. Since 416/547 = 0.76, 76% believe homosexual relations are always wrong. Likewise, you can check that 5% believe they are almost always wrong, 4% believe they are sometimes wrong, and 15% believe they are not wrong at all. For those who report being liberal, since 213/586 = 0.36, 36% believe homosexual relations are always wrong. Likewise, you can check that 5% believe they are almost always wrong, 9% believe they are sometimes wrong, and 50% believe they are not wrong at all. There seems to be a definite association between opinion about homosexuality and religious beliefs, with religious fundamentalists being more negative about homosexuality. Chapter 8 will show many other ways of analyzing data of this sort.

Bivariate Quantitative Data

When both variables are quantitative, a plot we've not yet discussed is helpful. Figure 3.19 shows an example using the software SPSS to plot data from 38 nations on fertility (the mean number of children per adult woman) and the percentage of the adult population using cell phones. (The data are shown later in the text in Table 9.13.) Here, values of cell-phone use are plotted on the horizontal axis, called the ***x-axis***, and values of fertility are plotted on the vertical axis, called the ***y-axis***. The values of the two variables for any particular observation form a point relative to these axes. To portray graphically the sample data, we plot the 38 observations as 38 points. For example, the point at the top left of the plot represents Pakistan, which had a fertility of 6.2 children per woman but cell-phone use of only 3.5%. This graphical plot is called a ***scatterplot***.

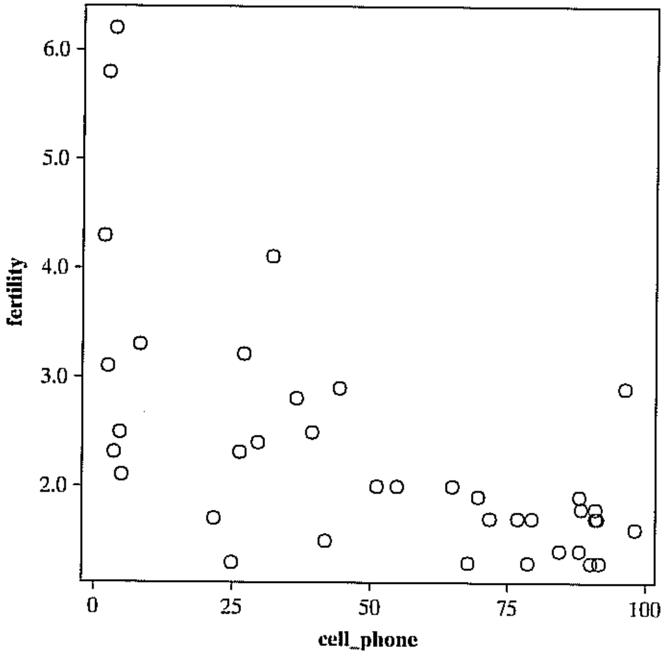

FIGURE 3.19: Scatterplot for Fertility and Percentage Using Cell Phones, for 38 Nations. The data are in Table 9.13 in Chapter 9.

The scatterplot shows a tendency for nations with higher cell-phone use to have lower levels of fertility. In Chapter 9 we'll learn about two ways to describe such as a trend. One way, called the ***correlation***, describes how strong the association is, in terms of how closely the data follow a *straight line trend*. For Figure 3.19, the correlation is -0.63. The negative value means that fertility tends to go *down* as cell-phone use goes *up*. By contrast, cell-phone use and GDP (gross domestic product, per capita) have a positive correlation of 0.83. As one goes up, the other also tends to go up.

The correlation takes values between -1 and $+1$. The larger it is in absolute value, that is, the farther from 0, the stronger the association. Cell-phone use is a bit more strongly associated with GDP than with fertility, because the correlation of 0.83 is larger in absolute value than the correlation of -0.63.

The second useful tool for describing the trend is ***regression analysis***. This provides a straight-line formula for predicting the value of the response variable from a given value of the explanatory variable. For Figure 3.19, this equation is

$$\text{Predicted fertility} = 3.4 - 0.02\,(\text{cell-phone use}).$$

For a country with no cell-phone use, the predicted fertility is $3.4 - 0.02(0) = 3.4$ children per mother. For a country with 100% of adults using cell phones, the predicted fertility is only $3.4 - 0.02(100) = 1.4$ children per mother.

Chapter 9 shows how to find the correlation and the regression line. Later chapters show how to extend the analysis to handle categorical as well as quantitative variables.

Analyzing More than Two Variables

This section has taken a quick look at analyzing associations between two variables. One important lesson from later in the text is that, *just because two variables have an association does not mean there is a causal connection*. For example, having more people in a nation using cell phones does not mean this is the reason the fertility rate is lower (for example, because people are talking on cell phones rather than doing what causes babies.) Perhaps high values on cell-phone use and low values on fertility are both a by-product of a nation being more economically advanced.

Most studies have *several* variables. The second half of this book (Chapters 10–16) shows how to conduct *multivariate* analyses. For example, to study what affects the number of good friends, we might want to simultaneously consider gender, age, whether married, educational level, whether attend religious services regularly, and whether live in urban or rural setting.

3.6 SAMPLE STATISTICS AND POPULATION PARAMETERS

Of the measures introduced in this chapter, the mean $\bar{y}$ is the most commonly reported measure of center and the standard deviation s is the most common measure of spread. We'll use them frequently in the rest of the text.

Since the values $\bar{y}$ and s depend on the sample selected, they vary in value from sample to sample. In this sense, they are variables. Their values are unknown before the sample is chosen. Once the sample is selected and they are computed, they become known sample statistics.

With inferential statistics, we shall distinguish between sample statistics and the corresponding measures for the population. Section 1.2 introduced the term *parameter* for a summary measure of the population. A statistic describes a sample, while a parameter describes the population from which the sample was taken. In this text, lowercase Greek letters usually denote population parameters and Roman letters denote the sample statistics.

Notation for Parameters

μ (Greek mu) and σ (Greek lowercase sigma) denote the mean and standard deviation of a variable for the population.

We call μ and σ the ***population mean*** and ***population standard deviation***. The population mean is the average of the observations for the entire population. The population standard deviation describes the variability of those observations about the population mean.

Whereas the statistics $\bar{y}$ and s are variables, with values depending on the sample chosen, the parameters μ and σ are constants. This is because μ and σ refer to just one particular group of observations, namely, the observations for the entire population. The parameter values are usually unknown, which is the reason for sampling and calculating sample statistics to estimate their values. Much of the rest of this text deals with ways of making inferences about unknown parameters (such as μ) using sample statistics (such as $\bar{y}$). Before studying these inferential methods, though, you need to learn some basic ideas of *probability*, which serves as the foundation for the methods. Probability is the subject of the next chapter.

3.7 CHAPTER SUMMARY

This chapter introduced ***descriptive statistics***—ways of *describing* data to summarize key characteristics of the data.

3.7.1 Overview of Tables and Graphs

- A ***frequency distribution*** summarizes the counts for possible values or intervals of values. A ***relative frequency*** distribution reports this information using percentages or proportions.
- A ***bar graph*** uses bars over possible values to portray a frequency distribution for a categorical variable. For a quantitative variable, a similar graphic is called a ***histogram***. It shows whether the distribution is approximately bell shaped, U shaped, skewed to the right (longer tail pointing to the right), or whatever.
- The ***stem-and-leaf plot*** is an alternative portrayal of data for a quantitative variable. It groups together observations having the same leading digit (stem), and shows also their final digit (leaf). For small samples, it displays the individual observations.
- The ***box plot*** portrays the quartiles, the extreme values, and any outliers. The box plot and the stem-and-leaf plot also can provide back-to-back comparisons of two groups.

Stem-and-leaf plots and box plots, simple as they are, are relatively recent innovations in statistics. They were introduced by the great statistician John Tukey (see Tukey 1977), who also introduced the terminology "software." See Cleveland (1994) and Tufte (2001) for other innovative ways to present data graphically.

3.7.2 Overview of Measures of Center

Measures of center describe the center of the data, in terms of a typical observation.

- The ***mean*** is the sum of the observations divided by the sample size. It is the center of gravity of the data.
- The ***median*** divides the ordered data set into two parts of equal numbers of observations, half below and half above that point.

- The lower quarter of the observations fall below the ***lower quartile***, and the upper quarter fall above the ***upper quartile***. These are the 25th and 75th ***percentiles***. The median is the 50th percentile. The quartiles and median split the data into four equal parts. They are less affected than the mean by outliers or extreme skew.
- The ***mode*** is the most commonly occurring value. It is valid for any type of data, though usually used with categorical data or discrete variables taking relatively few values.

3.7.3 Overview of Measures of Variability

Measures of variability describe the spread of the data.

- The ***range*** is the difference between the largest and smallest observations. The ***interquartile range*** is the range of the middle half of the data between the upper and lower quartiles. It is less affected by outliers.
- The ***variance*** averages the squared deviations about the mean. Its square root, the ***standard deviation***, is easier to interpret, describing a typical distance from the mean.
- The ***Empirical Rule*** states that for a bell-shaped distribution, about 68% of the observations fall within one standard deviation of the mean, about 95% fall within two standard deviations, and nearly all, if not all, fall within three standard deviations.

Table 3.9 summarizes the measures of center and variability. A ***statistic*** summarizes a sample. A ***parameter*** summarizes a population. ***Statistical inference*** uses statistics to make predictions about parameters.

TABLE 3.9: Summary of Measures of Center and Variability

Measure	Definition	Interpretation
Center		
Mean	$\bar{y} = \Sigma y_i/n$	Center of gravity
Median	Middle observation of ordered sample	50th percentile, splits sample into two equal parts
Mode	Most frequently occurring value	Most likely outcome, valid for all types of data
Variability		
Standard deviation	$s = \sqrt{\Sigma(y_i - \bar{y})^2/(n - 1)}$	Empirical Rule: If bell shaped, 68%, 95% within $s, 2s$ of $\bar{y}$
Range	Difference between largest and smallest observation	Greater with more variability
Interquartile range	Difference between upper quartile (75th percentile) and lower quartile (25th percentile)	Encompasses middle half of data

3.7.4 Overview of Bivariate Descriptive Statistics

Bivariate statistics are used to analyze data on two variables together.

- Many studies analyze how the outcome on a ***response variable*** depends on the value of an explanatory variable.

- For categorical variables, a ***contingency table*** shows the number of observations at the combinations of possible outcomes for the two variables.
- For quantitative variables, a ***scatterplot*** graphs the observations, showing a point for each observation. The response variable is plotted on the y-axis and the explanatory variable is plotted on the x-axis.
- For quantitative variables, the ***correlation*** describes the strength of straight-line association. It falls between −1 and +1 and indicates whether the response variable tends to increase (positive correlation) or decrease (negative correlation) as the explanatory variable increases.
- A ***regression analysis*** provides a straight-line formula for predicting the value of the response variable using the explanatory variable. We study correlation and regression in detail in Chapter 9.

PROBLEMS

Practicing the Basics

3.1. Table 3.10 shows the number (in millions) of the foreign-born population of the United States in 2004, by place of birth.

(a) Construct a relative frequency distribution.
(b) Sketch the data in a bar graph.
(c) Is "Place of birth" quantitative or categorical?
(d) Use whichever of the following measures is relevant for these data: mean, median, mode.

TABLE 3.10

Place of Birth	Number
Europe	4.7
Asia	8.7
Caribbean	3.3
Central America	12.9
South America	2.1
Other	2.6
Total	34.3

Source: Statistical Abstract of the United States, 2006.

3.2. According to www.adherents.com, in 2006 the number of followers of the world's five largest religions were 2.1 billion for Christianity, 1.3 billion for Islam, 0.9 billion for Hinduism, 0.4 billion for Confucianism, and 0.4 billion for Buddhism.

(a) Construct a relative frequency distribution.
(b) Sketch a bar graph.
(c) Can you find a mean, median, or mode for these data? If so, do so and interpret.

3.3. A teacher shows her class the scores on the midterm exam in the stem-and-leaf plot:

```
6 | 5 8 8
7 | 0 1 1 3 6 7 7 9
8 | 1 2 2 3 3 3 4 6 7 7 7 8 9
9 | 0 1 1 2 3 4 4 5 8
```

(a) Identify the number of students and the minimum and maximum scores.
(b) Sketch a histogram with four intervals.

3.4. According to the *2005 American Community Survey*, in 2005 the United States had 30.1 million households with one person, 37.0 million with two persons, 17.8 million with three persons, 15.3 million with four persons, and 10.9 million with five or more persons.

(a) Construct a relative frequency distribution.
(b) Sketch a histogram. What is its shape?
(c) Report and interpret the (i) median, (ii) mode of household size.

3.5. Copy the "2005 statewide crime" data file from the text Web site (www.stat.ufl.edu/~aa/social/data.html). Use the variable, murder rate (per 100,000 population). In this exercise, do not use the observation for D.C. Using software,

(a) Construct a relative frequency distribution.
(b) Construct a histogram. How would you describe the shape of the distribution?
(c) Construct a stem-and-leaf plot. How does this plot compare to the histogram in (b)?

3.6. The OECD (Organization for Economic Cooperation and Development) consists of advanced, industrialized countries that accept the principles of representative democracy and a free market economy. Table 3.11 shows UN data for OECD nations on several variables: gross domestic product (GDP, per capita in U.S. dollars), percent unemployed, a measure of inequality based on comparing wealth of the richest 10% to the poorest 10%, public expenditure on health (as a percentage of the GDP), the number of physicians per 100,000 people, carbon dioxide emissions (per capita, in metric tons), the percentage of seats in parliament held by women, and female economic activity as