Statistical Methods for the Social Sciences

Fifth Edition

Alan Agresti

University of Florida

Pearson

Chapter

3

Descriptive Statistics

CHAPTER OUTLINE

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.

Most importantly, the ***mean*** describes the center and the ***standard deviation*** describes the variability.

The final section introduces descriptive statistics that investigate, for a pair of variables, their ***association***—how certain values for one variable may tend to go with certain values of the other. For quantitative variables, the ***correlation*** describes the strength of the association, and ***regression analysis*** predicts the value of one variable from a value of the other variable.

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 variables, we list the categories and show the number of observations in each category. To make it easier to compare different categories, we also report proportions or percentages in the categories, also called *relative frequencies*. 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. The sum of the proportions equals 1.00. The sum of the percentages equals 100.

Example 3.1

Household Structure in the United States Table 3.1 lists the different types of households in the United States in 2015. Of 116.3 million households, for example, 23.3 million were a married couple with children, for a proportion of $23.3/116.3 = 0.20$.

A percentage is the proportion multiplied by 100. That is, the decimal place is moved two positions to the right. For example, since 0.20 is the proportion of families that are married couples with children, the percentage is $100(0.20) = 20\%$. Table 3.1

TABLE 3.1: U.S. Household Structure, 2015

Type of Family	Number (millions)	Proportion	Percentage (1970)
Married couple with children	23.3	0.20	20 (40)
Married couple, no children	33.7	0.29	29 (30)
Women living alone	17.4	0.15	15 (11)
Men living alone	14.0	0.12	12 (6)
Other family households	20.9	0.18	18 (11)
Other nonfamily households	7.0	0.06	6 (2)
Total	116.3	1.00	100 (100)

Source: U.S. Census Bureau; percentages from 1970 in parentheses.

also shows the percentages (in parentheses) from the year 1970. We see a substantial drop since 1970 in the relative number of married couples with children. ■

FREQUENCY DISTRIBUTIONS AND BAR GRAPHS: CATEGORICAL DATA

A table, such as Table 3.1, that lists the categories and their numbers of observations 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.

When the table shows the proportions or percentages instead of the numbers, it is called a ***relative frequency distribution***.

To more easily get a feel for the data, it is helpful to look at a graph of the frequency distribution. A ***bar graph*** has a rectangular bar drawn over each category. The height of the bar shows the frequency or 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.

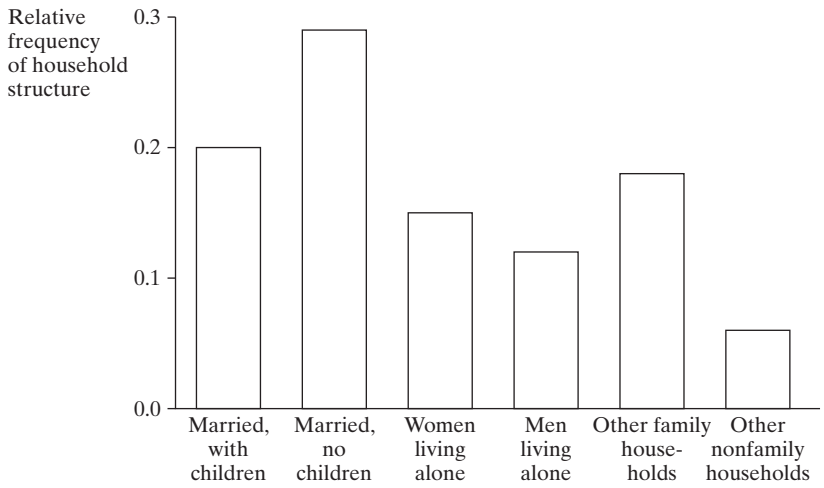


FIGURE 3.1: Bar Graph of Relative Frequency Distribution of U.S. Household Types

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. A bar graph is more precise than a 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 this.

Example 3.2

Statewide Violent Crime Rates Table 3.2 lists all 50 states in the United States and their 2015 violent crime rates. This rate measures the number of violent crimes in that state 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 $(12{,}000/2{,}300{,}000) \times 10{,}000 = 52$. Tables, graphs, and numerical measures help us absorb the information in these data.

TABLE 3.2: List of States with 2015 Violent Crime Rates Measured as Number of Violent Crimes per 10,000 Population

Alabama	43	Louisiana	52	Ohio	29
Alaska	64	Maine	13	Oklahoma	44
Arizona	42	Maryland	47	Oregon	25
Arkansas	46	Massachusetts	41	Pennsylvania	34
California	40	Michigan	45	Rhode Island	26
Colorado	31	Minnesota	23	South Carolina	51
Connecticut	26	Mississippi	27	South Dakota	32
Delaware	49	Missouri	43	Tennessee	59
Florida	47	Montana	25	Texas	41
Georgia	37	Nebraska	26	Utah	22
Hawaii	25	Nevada	60	Vermont	12
Idaho	22	New Hampshire	22	Virginia	20
Illinois	38	New Jersey	29	Washington	29
Indiana	36	New Mexico	61	West Virginia	30
Iowa	27	New York	39	Wisconsin	28
Kansas	34	North Carolina	34	Wyoming	21
Kentucky	21	North Dakota	27		

Source: `www.fbi.gov`; data are in `Crime` data file at text website.

To summarize the data with a frequency distribution, 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 {0–9, 10–19, 20–29, 30–39, 40–49, 50–59, 60–69}. Table 3.3 (page 32) shows that considerable variability exists in the violent crime rates.

Table 3.3 also shows the relative frequencies, using proportions and percentages. As with any summary method, we lose some information as the cost of achieving some clarity. The frequency distribution does not show the exact violent crime rates or identify which states have low or high rates. ■

The intervals of values in frequency distributions are usually of equal width. The width equals 10 in Table 3.3. The intervals should include all possible values of the variable. In addition, any possible value must fit into one and only one interval; that is, they should be ***mutually exclusive***.

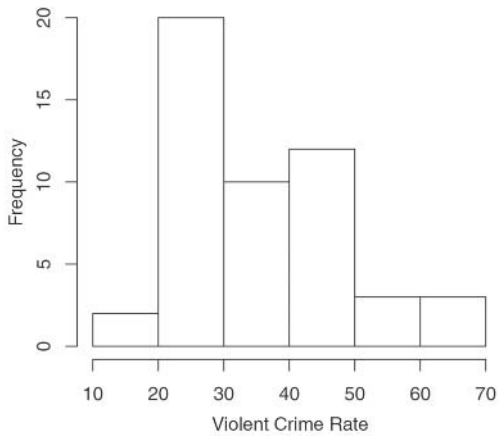
TABLE 3.3: Frequency Distribution and Relative Frequency Distribution for Violent Crime Rates

Violent Crime Rate	Frequency	Proportion	Percentage
0–9	0	0.00	0
10–19	2	0.04	4
20–29	20	0.40	40
30–39	10	0.20	20
40–49	12	0.24	24
50–59	3	0.06	6
60–69	3	0.06	6
Total	50	1.00	100.0

HISTOGRAMS

A graph of a 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, as constructed by R software.

FIGURE 3.2: Histogram of Frequencies for 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. 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.

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. Statistical software can automatically choose intervals for us and construct frequency distributions and histograms.

STEM-AND-LEAF PLOTS

Figure 3.3 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 first line, the stem of 1 and the leaves of 2 and 3 represent the violent crime rates 12 and 13. The plot arranges the leaves in order on each line, from smallest to largest.

FIGURE 3.3: Stem-and-Leaf Plot for Violent Crime Rate Data in Table 3.2

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

A stem-and-leaf plot conveys information similar to 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 3.3, the largest violent crime rate was 64, and the smallest was 12. It is not possible to determine these exact values from the histogram in Figure 3.2.

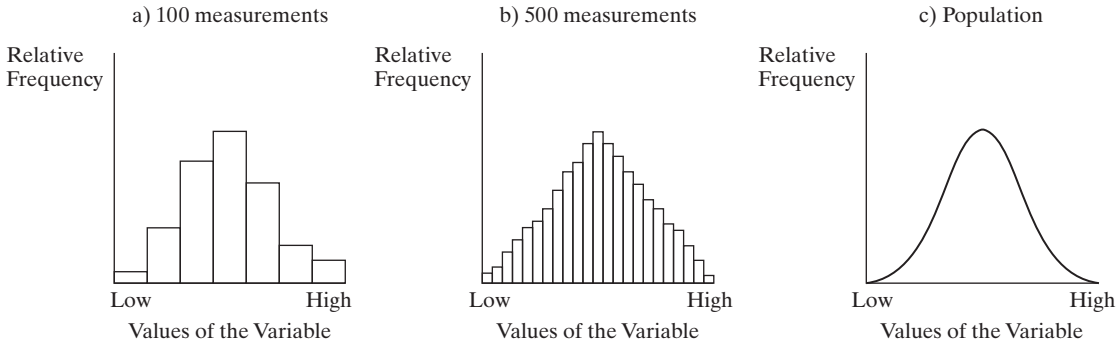
Stem-and-leaf plots are useful for quick portrayals of small data sets. As the sample size increases, you can accommodate the increase in leaves by splitting the stems. For instance, you can list each stem twice, putting leaves of 0 to 4 on one line and leaves of 5 to 9 on another. When a number has 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 a leaf of 6.

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.

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 3.4 shows two sample

FIGURE 3.4: Histograms for a Continuous Variable. We use smooth curves to represent population distributions for continuous variables.



histograms, one for a sample of size 100 and the second for 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.

THE SHAPE OF A DISTRIBUTION

Another way to describe a sample or a population distribution is by 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 3.5. 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.

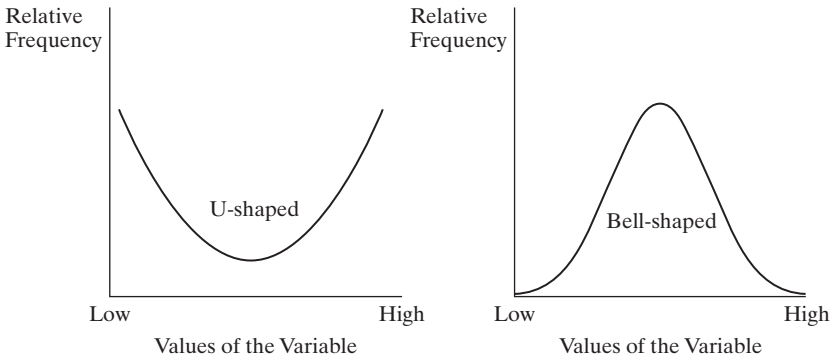


FIGURE 3.5: U-Shaped and Bell-Shaped Frequency Distributions

The distributions in Figure 3.5 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 3.6 illustrates this. The parts of the curve for the lowest values and the highest values are called the ***tails*** of the distribution. Often, as in Figure 3.6, 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.

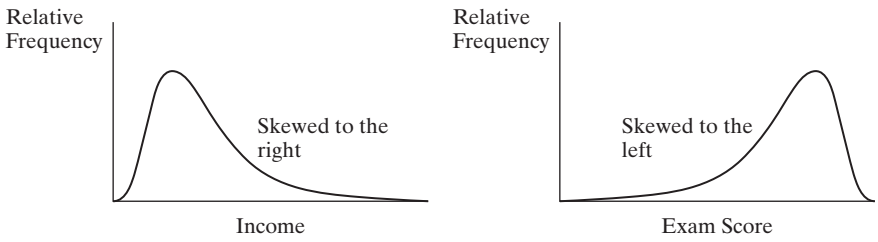


FIGURE 3.6: Skewed Frequency Distributions. The longer tail indicates the direction of skew.

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 violent crime rate for Southern states is 5.4 above the violent crime rate for Western states." We next present 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 3.3

Female Economic Activity in Europe and Middle East Table 3.4 shows an index of female economic activity in 2014 for the 10 largest countries (in population) of Western Europe and of the Middle East. The number specifies female employment as a percentage of male employment. In Italy, for instance, the number of females in the work force was 66% of the number of males in the work force.

TABLE 3.4: Female Employment, as a Percentage of Male Employment, in Western Europe and the Middle East

Western Europe		Middle East	
Country	Employment	Country	Employment
Belgium	79	Egypt	29
France	79	Iran	42
Germany	78	Iraq	19
Greece	68	Israel	81
Italy	66	Jordan	39
Netherlands	82	Saudi Arabia	4
Portugal	78	Syria	38
Spain	71	Turkey	34
Sweden	85	United Arab Emirates	49
UK	81	Yemen	40

Source: www.socialwatch.org.

For the 10 observations for Western Europe, the sum equals

$$79 + 79 + 78 + 68 + 66 + 82 + 78 + 71 + 85 + 81 = 767.$$

The mean equals $767/10 = 76.7$. By comparison, you can check that the mean for the 10 Middle Eastern countries equals $375/10 = 37.5$. Female economic activity tends to be considerably lower in the Middle East. ■

NOTATION FOR OBSERVATIONS, MEAN, AND SUMMATIONS

We use the following notation in formulas for the mean and 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 $\bar{y}$.

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 for the second, and so forth. For example, for female economic activity in Western Europe, $n = 10$, and the observations are $y_1 = 79, y_2 = 79, \ldots, y_{10} = 81$. The symbol $\bar{y}$ for the sample mean is read as "y-bar." 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 $\sum$ (upper case Greek letter sigma) represents the process of summing. For instance, $\sum y_i$ represents the sum $y_1 + y_2 + \cdots + y_n$. This symbol[1] 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 Western European data,

$$\sum y_i = y_1 + y_2 + \cdots + y_{10} = 79 + 79 + \cdots + 81 = 767.$$

Using this summation symbol, we have the shortened expression for the sample mean of n observations,

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

The summation operation is sometimes even further abbreviated as $\sum y$.

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, Muslim, Jewish, Other), the mean religion does not make sense, even though for convenience these levels may be coded in a data file by numbers.

- The mean can be highly influenced by an observation that falls well above or well below the bulk of the data, called an ***outlier***.

  Here is an example illustrating an outlier: The owner of Leonardo's Pizza reports that the mean annual income of full-time employees in the business is \$45,900. In fact, the annual incomes of the seven employees are \$15,400, \$15,600, \$15,900, \$16,400, \$16,400, \$16,600, and \$225,000. The \$225,000 income is the salary of the owner's son, who happens to be an employee. The value \$225,000 is an outlier. The mean computed for the other six observations alone equals \$16,050, quite different from the mean of \$45,900 including the outlier.

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

  In the Leonardo's Pizza example, the large observation \$225,000 results in an extreme skewness to the right of the income distribution. This skewness pulls the mean above six of the seven observations. This example shows that the mean is not always typical of the observations in the sample. The more highly skewed the distribution, the less typical the mean is of the data.

[1] You can also formally present the range of observations in the symbol, using $\sum_{i=1}^{n} y_i$ to represent summing y_i while letting i go from 1 to n.

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

  For example, Figure 3.7 shows that if we place an equal weight at each Middle Eastern observation on female economic activity from Table 3.4, then the line balances by placing a fulcrum at the point 37.5. The mean is the *center of gravity* (balance point) of the observations: 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.

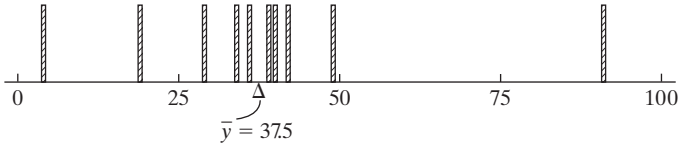


FIGURE 3.7: The Mean as the Center of Gravity, for Middle Eastern Data from Table 3.4. The line balances with a fulcrum at 37.5.

- 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 Western European observations have $n_1 = 10$ and $\bar{y}_1 = 76.70$. Canada, the United States, and Mexico have $n_2 = 3$ and values (83, 69, 56), for which $\bar{y}_2 = 69.33$. The overall mean economic activity for the 13 nations equals

$$\bar{y} = \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n_1 + n_2} = \frac{10(76.70) + 3(69.33)}{10 + 3} = \frac{(767 + 208)}{13} = \frac{975}{13} = 75.0.$$

The weighted average of 75.0 is closer to 76.7, the value for Western Europe, than to 69.3, the value for the three North American nations. This happens because more observations are from Western 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 or from highest to lowest.

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 of Leonardo's Pizza are

$15,400, $15,600, $15,900, $16,400, $16,400, $16,600, $225,000.

The median is the middle observation, $16,400. This is a more typical value for this sample than the sample mean of $45,900. 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 Western European nations are

$$66, 68, 71, 78, 78, 79, 79, 81, 82, 85.$$

Since $n = 10$ is even, the median is the midpoint between the two middle values, 78 and 79, which is $(78 + 79)/2 = 78.5$. This is close to the sample mean of 76.7, because this data set has no outliers.

The middle observation has the 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 $n = 10$, then $(n + 1)/2 = 5.5$, so the median is the midpoint between the fifth and sixth smallest observations.

Example 3.4

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 2014 American Community Survey taken by the U.S. Bureau of the Census. The possible responses form an ordinal scale. The population size was $n =$ 209 (in millions). The median score is the $(n + 1)/2 = (209 + 1)/2 = 105$th lowest. Now, 24 responses fall in the first category, $(24 + 62) = 86$ in the first two, $(24 + 62 + 35) = 121$ in the first three, and so forth. The 87th to 121st lowest scores fall in category 3, which therefore contains the 105th lowest, which is the median. The median response is "Some college, no degree." Equivalently, from the percentages in the last column of the table, $(11.5\% + 29.7\%) = 41.2\%$ fall in the first two categories and $(11.5\% + 29.7\% + 16.7\%) = 57.9\%$ 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	24	11.5
High school only	62	29.7
Some college, no degree	35	16.7
Associate's degree	21	10.0
Bachelor's degree	42	20.1
Master's degree	18	8.6
Doctorate or professional	7	3.3

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 3.5, the median and the mean are identical. To illustrate, the sample of observations 4, 5, 7, 9, and 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 longer tail relative to the median. See Figure 3.8.

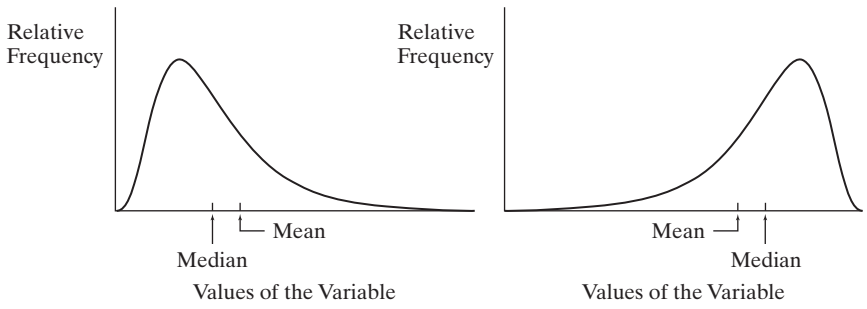


FIGURE 3.8: The Mean and the Median for Skewed Distributions. The mean is pulled in the direction of the longer tail.

The mean is larger than the median for distributions that are skewed to the right. For example, income distributions are often skewed to the right. Household income in the United States in 2015 had a mean of about $73,000 and a median of about $52,000 (U.S. Bureau of the Census).

The mean is smaller than the median for distributions that are skewed to the left. The distribution of grades on an exam may be skewed to the left when some students perform considerably poorer than the others. For example, suppose that an exam scored on a scale of 0 to 100 has a median of 88 and a mean of 80. 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 80.

- 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 Leonardo's Pizza employees have a median of $16,400 whether the largest observation is $20,000, $225,000, or $2,000,000.

MEDIAN COMPARED TO MEAN

The median is usually more appropriate than the mean when the distribution is very 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. 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, and 20 to the categories of highest degree, representing approximate number of years of education, we get a sample mean of 13.7.

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 the 2014 General Social Survey, summarizes the responses of the 53 females of age 18–22 to the question "How many sex partners have you had in the last 12 months?" Only six distinct responses occur, and 50.9% of those are 1. The median response is 1. For the sample mean, to sum the 52 observations we multiply each possible value by the frequency of its occurrence, and then add. That is,

$$\sum y_i = 11(0) + 27(1) + 6(2) + 5(3) + 3(4) + 1(5) = 71.$$

The sample mean response is

$$\bar{y} = \frac{\sum y_i}{n} = \frac{71}{53} = 1.34.$$

If the distribution of the 53 observations among these categories were (0, 27, 6, 5, 3, 12) (i.e., we shift the 11 responses from 0 to 5), then the median would still be 1, but the mean would shift to 2.38. 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 of Age 18–22

Response	Frequency	Percentage
0	11	20.8
1	27	50.9
2	6	11.3
3	5	9.4
4	3	5.7
5	1	1.9

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).

For binary (0, 1) data, proportion = mean

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 better than the mean in representing 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 over the median, because it uses the numerical values of all the observations.

THE MODE

Another measure, the *mode*, states the most frequent 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 3.6 on the number of sex partners in the last year, for instance, the mode is 1, since the frequency for that outcome is higher than the frequency for any other outcome. Here are some 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 grade given by 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 3.9 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 frequencies in the two extreme categories are much higher than those in the middle categories.

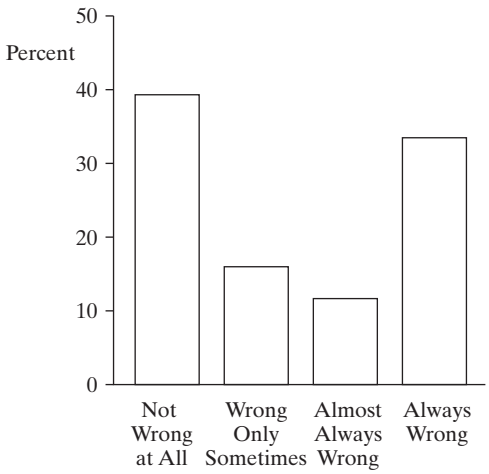


FIGURE 3.9: Bimodal Distribution for Opinion about Whether Abortion Is Wrong

- 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 the 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 3.10 illustrate this. 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.

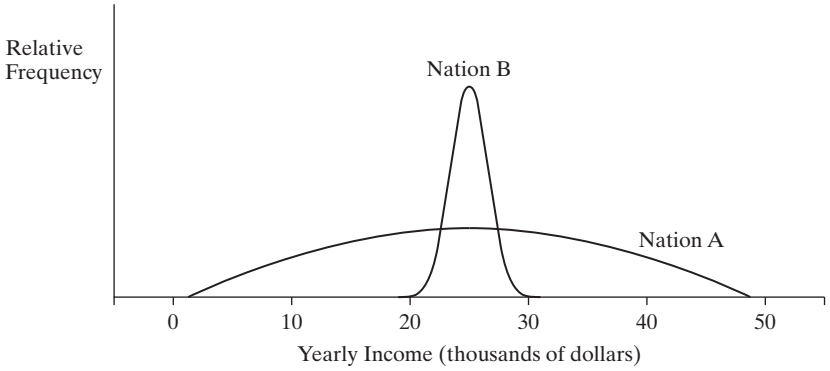


FIGURE 3.10:
Distributions with the Same Mean but Different Variability

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 3.10, 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 3.11 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.

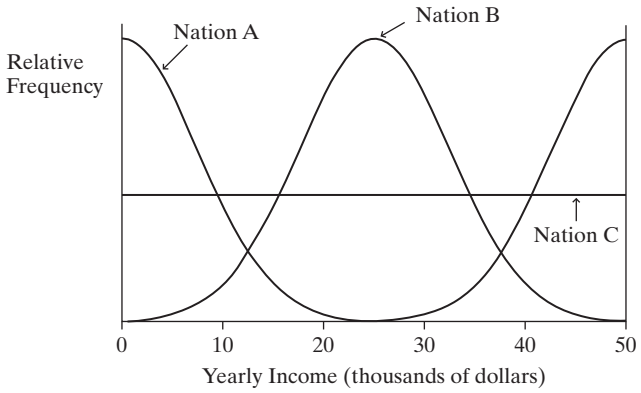


FIGURE 3.11:
Distributions with the Same Mean and Range, but Different Variability about the Mean

STANDARD DEVIATION

The most useful measure of variability is based on the *deviations* of the data from 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, $\sum(y_i - \bar{y})$, 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

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

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

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

The *variance* is approximately the 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 $\sum (y_i - \bar{y})^2$ in these formulas is called a ***sum of squares***. It represents squaring each deviation and then adding those squares. 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 simple: s is a sort of *typical distance* of an observation from the mean. So, *the larger the standard deviation, the greater the spread of the data*.

Example 3.5

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

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

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

For sample 1,

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

So, the variance is

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

and the standard deviation is $s = \sqrt{11.2} = 3.3$. 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. For example, for sample 2 the free software R finds

```
> quiz2 <- c(0, 0, 1, 9, 10, 10) # c COMBINES values listed
> sd(quiz2)              # sd is standard deviation function
[1] 5.138093
```

You should do the calculation yourself for a couple of small data sets to get a feel for what s 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, and 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.
- The reason for using $(n - 1)$, rather than n, in the denominator of s is technical. In Chapter 5, we'll see that doing this provides a better estimate of a corresponding parameter for the population. 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 about the population mean.
- 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: THE EMPIRICAL RULE

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 is applicable 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 frequency distributions seen in practice (i.e., *empirically*) are approximately bell shaped. Figure 3.12 is a graphical portrayal of the rule.

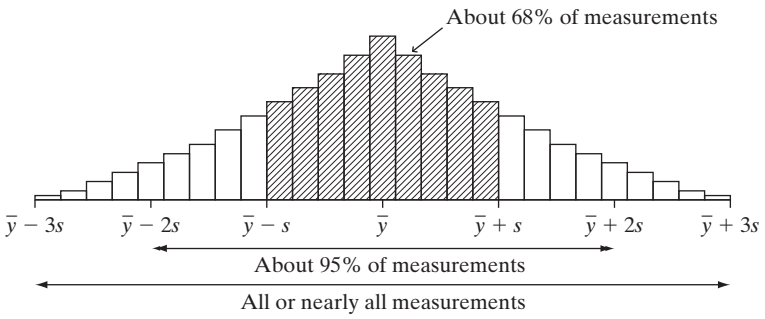


FIGURE 3.12: Empirical Rule: For Bell-Shaped Frequency Distributions, the Empirical Rule Specifies Approximate Percentages of Data within 1, 2, and 3 Standard Deviations of the Mean

Example 3.6

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 with mean about 500 and standard deviation about 100. Figure 3.13 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. ■

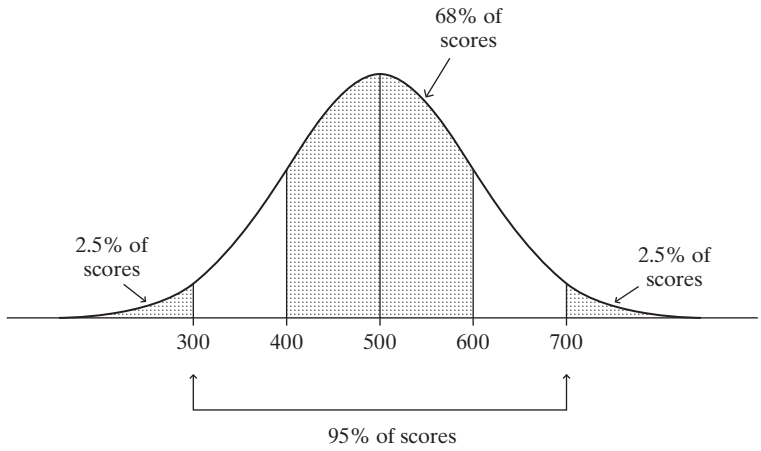


FIGURE 3.13: A Bell-Shaped Distribution of Scores for a Portion of the SAT, with Mean 500 and Standard Deviation 100

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 Empirical Rule does not apply 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 3.7

Familiarity with AIDS Victims A General Social Survey asked, "How many people have you known personally, either living or dead, who came down with AIDS?" Table 3.7 shows part of some software output for summarizing the 1598 responses on this variable. It indicates that 76% of the responses were 0.

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. 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. ■

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

Whenever the smallest or largest observation is less than a standard deviation from the mean, this is evidence of severe skew. Suppose that the first exam in your course, having potential scores between 0 and 100, has $\bar{y} = 86$ and $s = 15$. The upper bound of 100 is less than one standard deviation above the mean. The distribution is likely 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, for the incomes of the seven Leonardo's Pizza employees shown on page 36, $\bar{y} = \$45{,}900$ and $s = \$78{,}977$. When we remove the outlier, $\bar{y} = \$16{,}050$ and $s = \$489$.

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*.

Percentiles

> 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 $(100 - 50) = 50$%. In proportion terms, a percentile is called a **quantile**. The 50th percentile is the 0.50 quantile.

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 the percentile definition when we set $p = 25$ and $p = 75$. The quartiles together with the median split the distribution into four parts, each containing one-fourth of the observations. See Figure 3.14. 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.

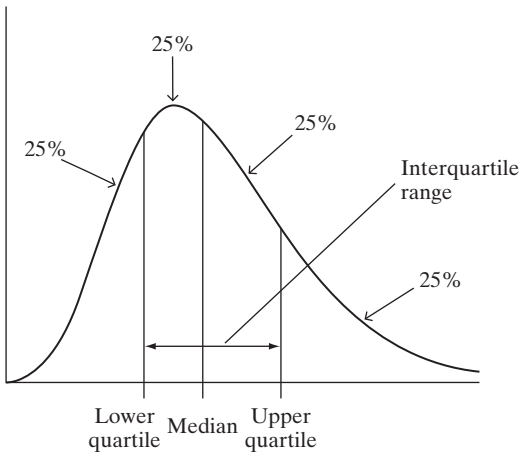


FIGURE 3.14: The Quartiles and the Median Split a Distribution into Four Equal Parts. The interquartile range describes the spread of the middle half of the distribution.

The median, the quartiles, and the maximum and minimum are five positions often used as a set to describe center and spread. Software can easily find these values as well as other percentiles. For instance, using R software we find $\bar{y}$ and s and then the five-number summary for the violent crime rates of Table 3.2, which the variable *violent* lists in the data file `Crime` at the text website:

```
> mean(violent); sd(violent)
[1] 34.9
[1] 12.43637
> summary(violent)
   Min.  1st Qu.  Median    Mean   3rd Qu.    Max.
   12.0     26.0    33.0    34.9      43.0    64.0
```

The lower and upper quartiles are labeled as "1st Qu." and "3rd Qu." In Stata, we use the `summarize` command to get $\bar{y}$, s, and the min and max.

```
. summarize violent
  Variable |   Obs      Mean    Std. Dev      Min      Max
   violent |    50      34.9    12.43637       12       64
```

We can also find the quartiles:

```
. tabstat violent, stats(p25 p50 p75)

    variable |       p25       p50       p75
     violent |        26        33        43
```

In summary, about a quarter of the states had violent crime rates (i) below 26, (ii) between 26 and 33, (iii) between 33 and 43, and (iv) above 43. The distance between the upper quartile and the median, $43 - 33 = 10$, exceeds the distance $33 - 26 = 7$ between the lower quartile and the median. This commonly happens when the distribution is skewed to the right.

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 just summarized by the five-number summary, the interquartile range IQR $= 43 - 26 = 17$. The middle half of the rates fall within a range of 17, whereas all rates fall within a range of $64 - 12 = 52$. 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, in 1990 the violent crime rates had quartiles of 33 and 77, giving an IQR of $77 - 33 = 44$. This indicates quite a bit more variability than in 2015, when IQR $= 17$.

An advantage of the IQR over the ordinary range or the standard deviation is that it is not sensitive to outliers. The violent crime rates ranged from 12 to 64, so the range was 52. When we include the observation for D.C., which was 130, the IQR changes only from 17 to 18. By contrast, the range changes from 52 to 118.

For bell-shaped distributions, the distance from the mean to either quartile is about two-thirds of a standard deviation. Then, IQR equals approximately $(4/3)s$.

BOX PLOTS: GRAPHING THE FIVE-NUMBER SUMMARY OF POSITIONS

The five-number summary consisting of (minimum, lower quartile, median, upper quartile, maximum) is the basis of a graphical display called[2] the ***box plot*** that summarizes center and 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*.

[2] Stem-and-leaf plots and box plots are relatively recent innovations, introduced by the statistician John Tukey (see Tukey, 1977), who also introduced the terminology "software."

These extend to the maximum and minimum, except for outliers, which are marked separately.

Figure 3.15 shows the box plot for the violent crime rates, including D.C., in the format provided with R software. The upper whisker and upper half of the central box are a bit 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.

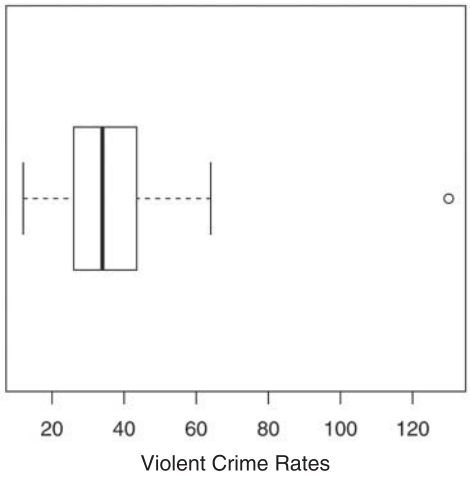


FIGURE 3.15: Box Plot of Violent Crime Rates of U.S. States. The outlier is the observation for D.C.

COMPARING GROUPS

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

Example 3.8

Comparing Canadian and U.S. Murder Rates Figure 3.16 (page 50) shows side-by-side box plots of murder rates (measured as the number of murders per 100,000 population) in a recent year for the 50 states in the United States 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 United States varied between 1.6 (Maine) and 20.3 (Louisiana). These side-by-side box plots reveal that the murder rates in the United States 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.

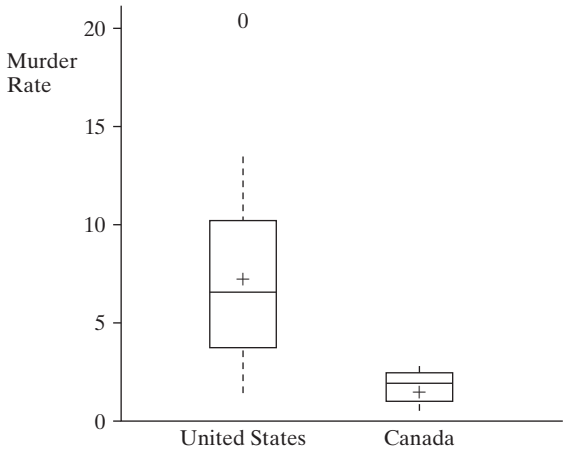


FIGURE 3.16: Box Plots for U.S. and Canadian Murder Rates

Figure 3.16 shows one outlier for the United States with a very high murder rate. This is the murder rate of 20.3 (for Louisiana). For these data, 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 value falls from the mean. For example, the U.S. murder rates shown in the box plot in Figure 3.16 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, 13.0 is 13.0/4.0 = 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 3.16, 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.

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 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. An association exists 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, Jewish, Muslim, Hindu, 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 the 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 and ethnic group is the explanatory 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 many other things in statistical science.

COMPARING TWO GROUPS: BIVARIATE CATEGORICAL AND QUANTITATIVE DATA

Chapter 7 presents 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 General Social Survey reports 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 highly 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 categories of the categorical variable on measures of the center (such as the mean and median) for the response variable. Graphs are also useful, such as side-by-side box plots.

BIVARIATE CATEGORICAL DATA

Chapter 8 presents 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 2014 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: Contingency Table for Religion and Opinion about Homosexual Relations

	Opinion about Homosexual Relations				
Religion	Always Wrong	Almost Always Wrong	Sometimes Wrong	Not Wrong at All	Total
Fundamentalist	262	10	19	87	378
Liberal	122	16	43	360	541

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. It's most 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 the *always wrong* category. For fundamentalists, since $262/378 = 0.69$, 69% believe homosexual relations are always wrong. For those who report being liberal, since $122/541 = 0.23$, 23% believe homosexual relations are always wrong. Likewise, you can check that the percentages responding *not wrong at all* were 23% for fundamentalists and 67% for liberals. There seems to be an appreciable association between opinion about homosexuality and religious beliefs, with religious fundamentalists being more negative about homosexuality. (For comparison, in the 1974 GSS the percentages in the *always wrong* category were 84% for fundamentalists and 47% for liberals, so the change in views over time has been considerable.) Chapter 8 shows many other ways of analyzing bivariate categorical data.

BIVARIATE QUANTITATIVE DATA

To illustrate methods that are useful when both variables are quantitative, we use the UN data file at the text website, partly shown in Table 3.9. The file has United Nations data from 2014 for 42 nations on per capita gross domestic product (GDP, in thousands of dollars), a human development index (HDI, which has components referring to life expectancy at birth, educational attainment, and income per capita), a gender inequality index (GII, a composite measure reflecting inequality in achievement between women and men in reproductive health, empowerment, and the labor market), fertility rate (number of births per woman), carbon dioxide emissions per capita (metric tons), a homicide rate (number of homicides per 100,000 people), prison population (per 100,000 people), and percent using the Internet.

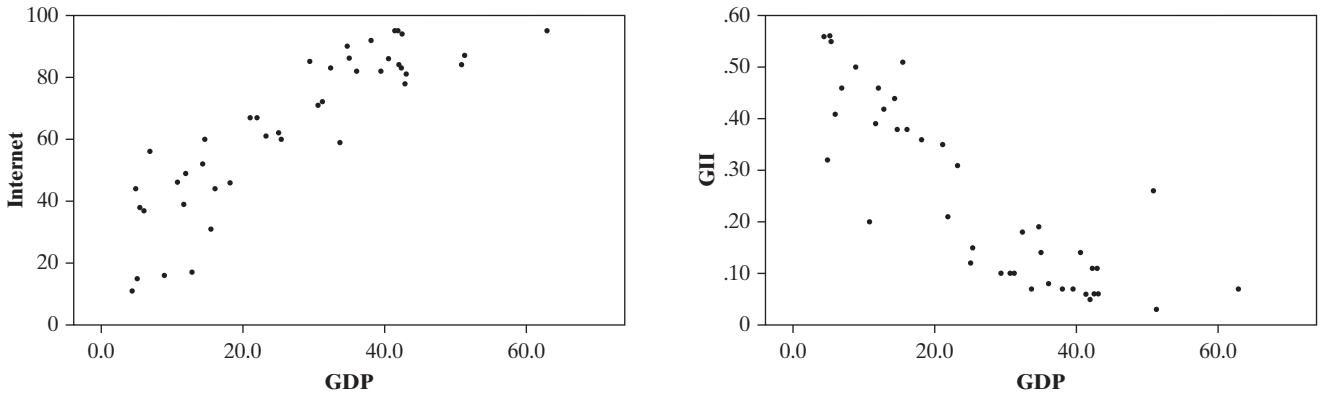
TABLE 3.9: National Data from UN Data File at Text Website

Nation	GDP	HDI	GII	Fertility	CO2	Homicide	Prison	Internet
Algeria	12.8	0.72	0.42	2.8	3.2	0.8	162	17
Argentina	14.7	0.81	0.38	2.2	4.7	5.5	147	60
Australia	42.3	0.93	0.11	1.9	16.5	1.1	130	83
Austria	43.1	0.88	0.06	1.4	7.8	0.8	98	81
Belgium	39.5	0.88	0.07	1.8	8.8	1.8	108	82
Brazil	14.3	0.74	0.44	1.8	2.2	21.8	274	52
Canada	40.6	0.90	0.14	1.6	14.1	1.5	118	86
...								
UK	34.7	0.89	0.19	1.9	7.1	1.2	148	90
US	50.9	0.91	0.26	1.9	17.0	4.7	716	84
Vietnam	4.9	0.64	0.32	1.7	2.0	1.6	145	44

Source: http://hdr.undp.org/en/data and http://data.worldbank.org; complete data file UN ($n = 42$) is at text website.

Figure 3.17 is an example of a type of graphical plot, called a ***scatterplot***, that portrays bivariate relations between quantitative variables. It plots data on percent using the Internet and gross domestic product. Here, values of GDP are plotted on the horizontal axis, called the ***x-axis***, and values of Internet use 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. The figure plots the 42 observations as 42 points. For example, the point at the highest level on GDP represents Norway, which had a GDP of 62.9 and Internet use of 95 percent. The scatterplot shows a tendency for nations with higher GDP to have higher levels of Internet use.

FIGURE 3.17: Scatterplots for GDP as Predictor of Internet Use and of GII, for 42 Nations



In Chapter 9, we'll learn about two ways to describe such a trend. One way to describe the trend, called the ***correlation***, describes how strong the association is, in terms of how closely the data follow a *straight-line trend*. For Figure 3.17, the correlation is 0.88. The positive value means that Internet use tends to go *up* as GDP goes *up*. By contrast, Figure 3.17 also shows a scatterplot for GDP and GII. Those variables have a negative correlation of −0.85. As GDP goes up, GII tends to go down. 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. For example, GDP is more strongly associated with Internet use and with GII than it is with fertility, because correlations of 0.88 and −0.85 are larger in absolute value than the correlation of −0.49 between GDP and fertility.

The second useful tool for describing the trend is ***regression analysis***. This method treats one variable, usually denoted by y, as the response variable, and the other variable, usually denoted by x, as the explanatory variable. It provides a straight-line formula for predicting the value of y from a given value of x. For the data from Table 3.9 on $y =$ fertility rate and $x =$ GDP, this equation is

$$\text{Predicted fertility} = 2.714 - 0.025(\text{GDP}).$$

For a country with GDP = 4.4 (the lowest value in this sample), the predicted fertility rate is $2.714 - 0.025(4.4) = 2.6$ births per woman. For a country with GDP = 62.9 (the highest value in this sample), the predicted fertility rate is $2.714 - 0.025(62.9) = 1.1$ births per woman.

Chapter 9 shows how to find the correlation and the regression line. It is simple with software, as shown in Table 3.10 using R with variables from the data file UN at the text website. Later chapters show how to extend the analysis to handle categorical as well as quantitative variables.

TABLE 3.10: Using R Software for a Scatterplot, Correlation, and Regression Line

```
> UN <- read.table("http://www.stat.ufl.edu/~aa/smss/data/UN.dat",
+                  header=TRUE)
> attach(UN)
> plot(GDP, Fertility)  # requests scatterplot
> cor(GDP, Fertility); cor(GDP, Internet); cor(GDP, GII)
[1] -0.4861589
[1]  0.8771987
[1] -0.8506693

> lm(Fertility ~ GDP)  # lm is short for "linear model"
Coefficients:
(Intercept)          GDP
    2.71401     -0.02519
```

ANALYZING MORE THAN TWO VARIABLES

This section has introduced 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, the correlation for Table 3.9 between the Internet use and the fertility rate is −0.48. But having more people using the Internet need not be the reason the fertility rate tends to be lower (e.g., because people are on the Internet rather than doing what causes babies). Perhaps high values on Internet 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 is associated with 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 used 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 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, lower case Greek letters usually denote population parameters and Roman letters denote the sample statistics.

Notation for Mean and Standard Deviation Parameters

> Greek letters denote parameters. For example, μ (mu) and σ (sigma) denote the population mean and standard deviation of a variable.

We call μ and σ the ***population mean*** and ***population standard deviation***, respectively. 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 computing sample statistics to estimate their values. Much of the rest of this text deals with ways of making inferences about 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.

OVERVIEW OF TABLES AND GRAPHS

- A ***frequency distribution*** summarizes numbers of observations for possible values or intervals of values of a variable.
- For a quantitative variable, a ***histogram*** uses bars over possible values or intervals of values to portray a frequency distribution. It shows shape—such as whether the distribution is approximately bell shaped or skewed to the right (longer tail pointing to the right) or to the left.
- The ***box plot*** portrays the quartiles, the extreme values, and any outliers.

Cook (2014) and Tufte (2001) showed other innovative ways to present data graphically.

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. These ***measures of position***, portrayed with extreme values in ***box plots***, are less affected than the mean by outliers or extreme skew.

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 of the mean, and nearly all, if not all, fall within three standard deviations of the mean.

Table 3.11 summarizes 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 3.11: Summary of Measures of Center and Variability

Measure	Definition	Interpretation
Center		
Mean	$\bar{y} = \sum y_i/n$	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	$s = \sqrt{\sum(y_i - \bar{y})^2/(n-1)}$	Empirical Rule: If bell shaped, 68%, 95% within s, $2s$ of $\bar{y}$
Range	Largest − smallest observation	Greater with more variability
Interquartile range	Upper quartile (75th percentile) − lower quartile (25th percentile)	Encompasses middle half of data

OVERVIEW OF BIVARIATE DESCRIPTIVE STATISTICS

Bivariate statistics summarize data on two variables together, to analyze the ***association*** between them.

- 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. It shows a point for each observation, plotting the response variable on the y-axis and the explanatory variable 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 line*** is a straight-line formula for predicting the response variable using the explanatory variable.

Exercises

Practicing the Basics

3.1. Table 3.12 shows the number (in millions) of the foreign-born population of the United States, 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 3.12

Place of Birth	Number
Europe	4.5
Asia	10.1
Caribbean	3.6
Central America	14.4
South America	2.4
Other	2.6
Total	37.6

Source: *Statistical Abstract of the United States, 2012.*

3.2. According to the 2013–2014 edition of *The World Factbook*, the number of followers of the world's four largest religions was 2.2 billion for Christianity, 1.6 billion for Islam, 1.0 billion for Hinduism, and 0.5 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 corresponding histogram with four intervals.

3.4. According to the *2015 American Community Survey*, in 2012 the United States had 30.1 million households with one person, 37.1 million with two persons, 17.8 million with three persons, 15.0 million with four persons, and 10.4 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. Create a data file with your software for the `Crime` data file from the text website. Use the variable *murder*, which is the murder rate (per 100,000 population). 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)?