

CHAPTER 2

DESCRIBING DATA

2.1 OVERVIEW

The starting point for data analysis is a data table (often referred to as a data set) which contains the measured or collected data values represented as numbers or text. The data in these tables are called *raw* before they have been transformed or modified. These data values can be measurements of a patient's weight (such as 150 lb, 175 lb, and so on) or they can be different industrial sectors (such as the "telecommunications industry," "energy industry," and so on) used to categorize a company. A data table lists the different items over which the data has been collected or measured, such as different patients or specific companies. In these tables, information considered interesting is shown for different attributes. The individual items are usually shown as rows in a data table and the different attributes shown as columns. This chapter examines ways in which individual attributes can be described and summarized: the scales on which they are measured, how to describe their center as well as the variation using *descriptive statistical* approaches, and how to make statements about these attributes using *inferential statistical* methods, such as confidence intervals or hypothesis tests.

2.2 OBSERVATIONS AND VARIABLES

All disciplines collect data about items that are important to that field. Medical researchers collect data on patients, the automotive industry on cars, and retail companies on transactions. These items are organized into a table for data analysis where each row, referred to as an *observation*, contains information about the specific item the row represents. For example, a data table about cars may contain many observations on different types of cars. Data tables also contain information about the car, for example, the car's weight, the number of cylinders, the fuel efficiency, and so on. When an attribute is thought of as a set of values describing some aspect across all observations, it is called a *variable*. An example of a table describing different attributes of cars is shown in Table 2.1 from Bache & Lichman (2013). Each row of the table describes an observation (a specific car) and each column describes a variable (a specific attribute of a car). In this example, there are five observations ("Chevrolet Chevelle Malibu," "Buick Skylark 320," "Plymouth Satellite," "AMC Rebel SST," "Ford Torino") and these observations are described using nine variables: *Name*, *MPG*, *Cylinders*, *Displacement*, *Horsepower*, *Weight*, *Acceleration*, *Model year*, and *Origin*. (It should be noted that throughout the book variable names in the text will be italicized.)

A generalized version of the data table is shown in Table 2.2, since a table can represent any number of observations described over multiple variables. This table describes a series of observations (from o_1 to o_n) where each observation is described using a series of variables (from x_1 to x_p). A value is provided for each variable of each observation. For example, the value of the first observation for the first variable is x_{11} , the value for the second observation's first variable is x_{21} , and so on. Throughout the book we will explore different mathematical operations that make use of this generalized form of a data table.

The most common way of looking at data is through a spreadsheet, where the raw data is displayed as rows of observations and columns of variables. This type of visualization is helpful in reviewing the raw data; however, the table can be overwhelming when it contains more than a handful of observations or variables. Sorting the table based on one or more variables is useful for organizing the data; however, it is difficult to identify trends or relationships by looking at the raw data alone. An example of a spreadsheet of different cars is shown in Figure 2.1.

Prior to performing data analysis or data mining, it is essential to understand the data table and an important first step is to understand in detail the individual variables. Many data analysis techniques have restrictions on

TABLE 2.1 Data Table Showing Five Car Records Described by Nine Variables

Name	MPG	Cylinders	Displacement	Horsepower	Weight	Acceleration	Model Year	Origin
Chevrolet Chevelle Malibu	18	8	307	130	3504	12	70	America
Buick Skylark 320	15	8	350	165	3693	11.5	70	America
Plymouth Satellite	18	8	318	150	3436	11	70	America
AMC Rebel SST	16	8	304	150	3433	12	70	America
Ford Torino	17	8	302	140	3449	10.5	70	America

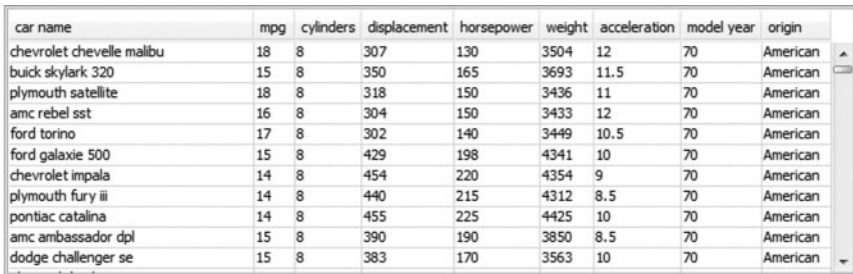
TABLE 2.2 Generalized Form of a Data Table

		Variables				
Observations		x_1	x_2	x_3	\dots	x_p
	o_1	x_{11}	x_{12}	x_{13}	\dots	x_{1p}
	o_2	x_{21}	x_{22}	x_{23}	\dots	x_{2p}
	o_3	x_{31}	x_{32}	x_{33}	\dots	x_{3p}
	\dots	\dots	\dots	\dots	\dots	\dots
	o_n	x_{n1}	x_{n2}	x_{n3}	\dots	x_{np}

the types of variables that they are able to process. As a result, knowing the types of variables allow these techniques to be eliminated from consideration or the data must be transformed into a form appropriate for analysis. In addition, certain characteristics of the variables have implications in terms of how the results of the analysis will be interpreted.

2.3 TYPES OF VARIABLES

Each of the variables within a data table can be examined in different ways. A useful initial categorization is to define each variable based on the type of values the variable has. For example, does the variable contain a fixed number of distinct values (*discrete* variable) or could it take any numeric value (*continuous* variable)? Using the examples from Section 2.1, an *industrial sector* variable whose values can be “telecommunication industry,” “retail industry,” and so on is an example of a discrete variable since there are a finite number of possible values. A patient’s *weight* is an example of a continuous variable since any measured value, such as 153.2 lb, 98.2 lb, is possible within its range. Continuous variables may have an infinite number of values.



car name	mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin
chevrolet chevelle malibu	18	8	307	130	3504	12	70	American
buick skylark 320	15	8	350	165	3693	11.5	70	American
plymouth satellite	18	8	318	150	3436	11	70	American
amc rebel sst	16	8	304	150	3433	12	70	American
ford torino	17	8	302	140	3449	10.5	70	American
ford galaxie 500	15	8	429	198	4341	10	70	American
chevrolet impala	14	8	454	220	4354	9	70	American
plymouth fury iii	14	8	440	215	4312	8.5	70	American
pontiac catalina	14	8	455	225	4425	10	70	American
amc ambassador dpl	15	8	390	190	3850	8.5	70	American
dodge challenger se	15	8	383	170	3563	10	70	American

FIGURE 2.1 Spreadsheet showing a sample of car observation.

Variables may also be classified according to the *scale* on which they are measured. Scales help us understand the precision of an individual variable and are used to make choices about data visualizations as well as methods of analysis.

A *nominal scale* describes a variable with a limited number of different values that cannot be ordered. For example, a variable *Industry* would be nominal if it had categorical values such as “financial,” “engineering,” or “retail.” Since the values merely assign an observation to a particular category, the order of these values has no meaning.

An *ordinal scale* describes a variable whose values can be ordered or ranked. As with the nominal scale, values are assigned to a fixed number of categories. For example, a scale where the only values are “low,” “medium,” and “high” tells us that “high” is larger than “medium” and “medium” is larger than “low.” However, although the values are ordered, it is impossible to determine the magnitude of the difference between the values. You cannot compare the difference between “high” and “medium” with the difference between “medium” and “low.”

An *interval scale* describes values where the interval between values can be compared. For example, when looking at three data values measured on the Fahrenheit scale—5°F, 10°F, 15°F—the differences between the values 5 and 10, and between 10 and 15 are both 5°. Because the intervals between values in the scale share the same unit of measurement, they can be meaningfully compared. However, because the scale lacks a meaningful zero, the ratios of the values cannot be compared. Doubling a value does not imply a doubling of the actual measurement. For example, 10°F is not twice as hot as 5°F.

A *ratio scale* describes variables where both intervals between values and ratios of values can be compared. An example of a ratio scale is a bank account balance whose possible values are \$5, \$10, and \$15. The difference between each pair is \$5; and \$10 is twice as much as \$5. Scales for which it is possible to take ratios of values are defined as having a natural zero.

A variable is referred to as *dichotomous* if it can contain only two values. For example, the values of a variable *Gender* may only be “male” or “female.” A *binary* variable is a widely used dichotomous variable with values 0 or 1. For example, a variable *Purchase* may indicate whether a customer bought a particular product using 0 to indicate that a customer did not buy and 1 to indicate that they did buy; or a variable *Fuel Efficiency* may use 0 to represent low efficiency vehicles and 1 to represent high efficiency vehicles. Binary variables are often used in data analysis because they provide a convenient numeric representation for many different types of discrete data and are discussed in detail throughout the book.

Certain types of variables are not used directly in data analysis, but may be helpful for preparing data tables or interpreting the results of the analysis. Sometimes a variable is used to identify each observation in a data table, and will have unique values across the observations. For example, a data table describing different cable television subscribers may include a *customer reference number* variable for each customer. You would never use this variable in data analysis since the values are intended only to provide a link to the individual customers. The analysis of the cable television subscription data may identify a subset of subscribers that are responsible for a disproportionate amount of the company's profit. Including a unique identifier provides a reference to detailed customer information not included in the data table used in the analysis. A variable may also have identical values across the observations. For example, a variable *Calibration* may define the value of an initial setting for a machine used to generate a particular measurement and this value may be the same for all observations. This information, although not used directly in the analysis, is retained both to understand how the data was generated (i.e., what was the calibration setting) and to assess the data for accuracy when it is merged from different sources. In merging data tables generated from two sensors, if the data was generated using different calibration settings then either the two tables cannot be merged or the calibration setting needs to be included to indicate the difference in how the data was measured.

Annotations of variables are another level of detail to consider. They provide important additional information that give insight about the context of the data: Is the variable a count or a fraction? A time or a date? A financial term? A value derived from a mathematical operation on other variables? The units of measurement are useful when presenting the results and are critical for interpreting the data and understanding how the units should align or which transformations apply when data tables are merged from different sources.

In Chapter 6, we further categorize variables (*independent variables* and *response variables*) by the roles they play in the mathematical models generated from data tables.

2.4 CENTRAL TENDENCY

2.4.1 Overview

Of the various ways in which a variable can be summarized, one of the most important is the value used to characterize the center of the set of values it contains. It is useful to quantify the middle or central location of a variable, such as its average, around which many of the observations'

values for that variable lie. There are several approaches to calculating this value and which is used can depend on the classification of the variable. The following sections describe some common descriptive statistical approaches for calculating the central location: the *mode*, the *median*, and the *mean*.

2.4.2 Mode

The *mode* is the most commonly reported value for a particular variable. The mode calculation is illustrated using the following variable whose values (after being ordered from low to high) are

3, 4, 5, 6, 7, 7, 7, 8, 8, 9

The mode would be the value 7 since there are three occurrences of 7 (more than any other value). The mode is a useful indication of the central tendency of a variable, since the most frequently occurring value is often toward the center of the variable's range.

When there is more than one value with the same (and highest) number of occurrences, either all values are reported or a midpoint is selected. For example, for the following values, both 7 and 8 are reported three times:

3, 4, 5, 6, 7, 7, 7, 8, 8, 8, 9

The mode may be reported as {7, 8} or 7.5.

Mode provides the only measure of central tendency for variables measured on a nominal scale; however, the mode can also be calculated for variables measured on the ordinal, interval, and ratio scales.

2.4.3 Median

The *median* is the middle value of a variable, once it has been sorted from low to high. The following set of values for a variable will be used to illustrate:

3, 4, 7, 2, 3, 7, 4, 2, 4, 7, 4

Before identifying the median, the values must be sorted:

2, 2, 3, 3, 4, 4, 4, 4, 7, 7, 7

There are 11 values and therefore the sixth value (five values above and five values below) is selected as the median value, which is 4:

2, 2, 3, 3, 4, 4, 4, 4, 7, 7, 7

For variables with an even number of values, the average of the two values closest to the middle is selected (sum the two values and divide by 2).

The median can be calculated for variables measured on the ordinal, interval, and ratio scales and is often the best indication of central tendency for variables measured on the ordinal scale. It is also a good indication of the central value for a variable measured on the interval or ratio scales since, unlike the mean, it will not be distorted by extreme values.

2.4.4 Mean

The *mean*—commonly referred to as the average—is the most commonly used summary of central tendency for variables measured on the interval or ratio scales. It is defined as the sum of all the values divided by the number of values. For example, for the following set of values:

3, 4, 5, 7, 7, 8, 9, 9, 9

The sum of all nine values is $(3 + 4 + 5 + 7 + 7 + 8 + 9 + 9 + 9)$ or 61. The sum divided by the number of values is $61 \div 9$ or 6.78.

For a variable representing a subset of all possible observations (x), the mean is commonly referred to as \bar{x} . The formula for calculating a mean, where n is the number of observations and x_i is the individual values, is usually written:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

The notation $\sum_{i=1}^n$ is used to describe the operation of summing all values of x from the first value ($i = 1$) to the last value ($i = n$), that is $x_1 + x_2 + \cdots + x_n$.

2.5 DISTRIBUTION OF THE DATA

2.5.1 Overview

While the central location is a single value that characterizes an individual variable's data values, it provides no insight into the variation of the data or, in other words, how the different values are distributed around this

location. The frequency distribution, which is based on a simple count of how many times a value occurs, is often a starting point for the analysis of variation. Understanding the frequency distribution is the focus of the following section and can be performed using simple data visualizations and calculated metrics. As you will see later, the frequency distribution also plays a role in selecting which data analysis approaches to adopt.

2.5.2 Bar Charts and Frequency Histograms

Visualization is an aid to understanding the distribution of data: the range of values, the shape created when the values are plotted, and the values called *outliers* that are found by themselves at the extremes of the range of values. A handful of charts can help to understand the frequency distribution of an individual variable. For a variable measured on a nominal scale, a *bar chart* can be used to display the relative frequencies for the different values. To illustrate, the *Origin* variable from the auto-MPG data table (partially shown in Table 2.2) has three possible values: “America,” “Europe,” and “Asia.” The first step is to count the number of observations in the data table corresponding to each of these values. Out of the 393 observations in the data table, there are 244 observations where the *Origin* is “America,” 79 where it is “Asia,” and 70 where it is “Europe.” In a bar chart, each bar represents a value and the height of the bars is proportional to the frequency, as shown in Figure 2.2.

For nominal variables, the ordering of the *x*-axis is arbitrary; however, they are often ordered alphabetically or based on the frequency value. The

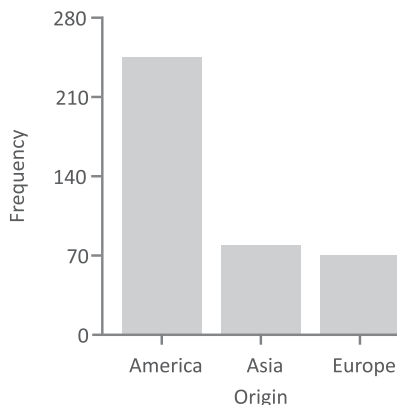


FIGURE 2.2 Bar chart for the *Origin* variable from the auto-MPG data table.

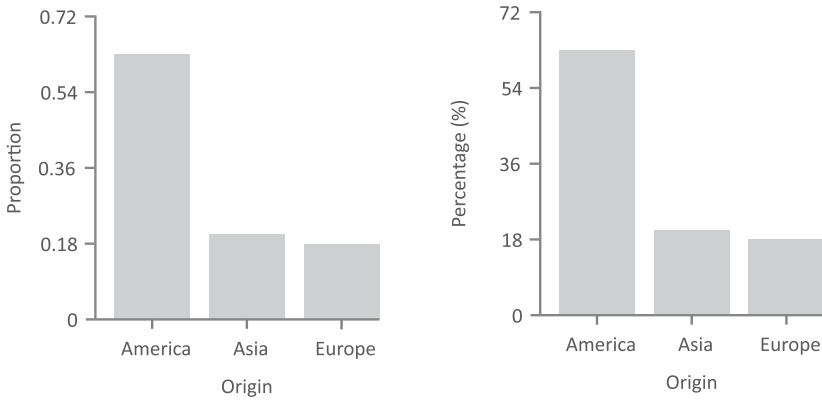


FIGURE 2.3 Bar charts for the *Origin* variables from the auto-MPG data table showing the proportion and percentage.

y-axis which measures frequency can also be replaced by values representing the proportion or percentage of the overall number of observations (replacing the frequency value), as shown in Figure 2.3.

For variables measured on an ordinal scale containing a small number of values, a bar chart can also be used to understand the relative frequencies of the different values. Figure 2.4 shows a bar chart for the variable *PLT* (number of mother’s previous premature labors) where there are four possible values: 1, 2, 3, and 4. The bar chart represents the number of values for each of these categories. In this example you can see that most of the observations fall into the “1” category with smaller numbers in the other categories. You can also see that the number of observations decreases as the values increase.

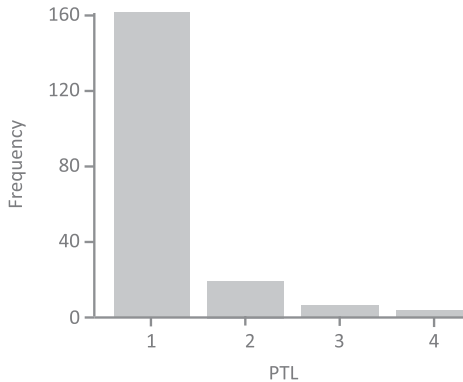


FIGURE 2.4 Bar chart for a variable measured on an ordinal scale, *PLT*.

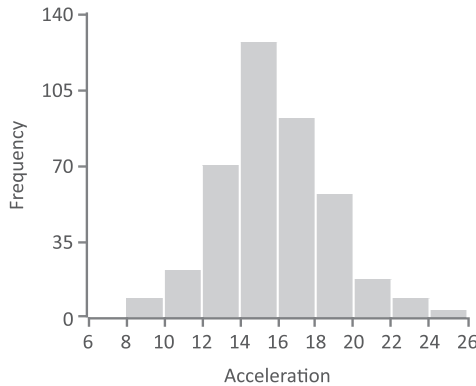


FIGURE 2.5 Frequency histogram for the variable “Acceleration.”

The frequency histogram is useful for variables with an ordered scale—ordinal, interval, or ratio—that contain a larger number of values. As with the bar chart, each variable is divided into a series of groups based on the data values and displayed as bars whose heights are proportional to the number of observations within each group. However, the criteria for inclusion within a single bar is a specific range of values. To illustrate, a frequency histogram is shown in Figure 2.5 displaying a frequency distribution for a variable *Acceleration*. The variable has been grouped into a series of ranges from 6 to 8, 8 to 10, 10 to 12, and so on. Since we will need to assign observations that fall on the range boundaries to only one category, we will assign a value to a group where its value is greater than or equal to the lower extreme and less than the upper extreme. For example, an *Acceleration* value of 10 will be categorized into the range 10–12. The number of observations that fall within each range is then determined. In this case, there are six observations that fall into the range 6–8, 22 observations that fall into the range 8–10, and so on. The ranges are ordered from low to high and plotted along the *x*-axis. The height of each histogram bar corresponds to the number of observations for each of the ranges. The histogram in Figure 2.5 indicates that the majority of the observations are grouped in the middle of the distribution between 12 and 20 and there are relatively fewer observations at the extreme values. It is usual to display between 5 and 10 groups in a frequency histogram using boundary values that are easy to interpret.

The frequency histogram helps to understand the shape of the frequency distribution. Figure 2.6 illustrates a number of commonly encountered frequency distributions. The first histogram illustrates a variable where, as the values increase, the number of observations in each group remains

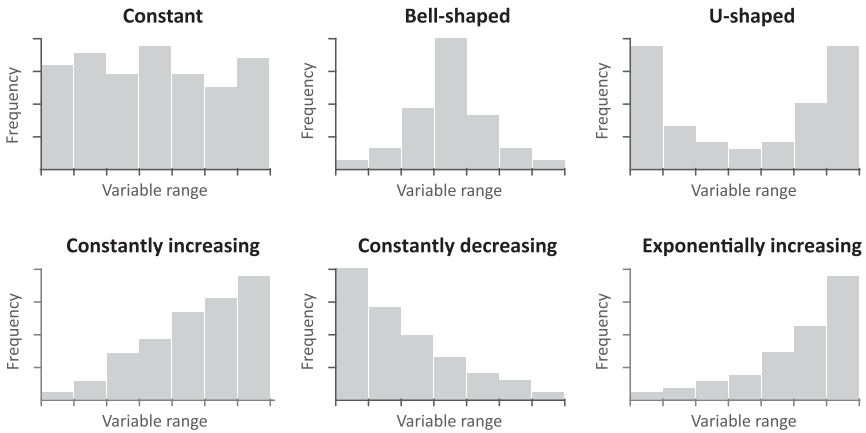


FIGURE 2.6 Examples of frequency distributions.

constant. The second histogram is of a distribution where most of the observations are centered around the mean value, with far fewer observations at the extremes, and with the distribution tapering off toward the extremes. The symmetrical shape of this distribution is often identified as a bell shape and described as a *normal* distribution. It is very common for variables to have a normal distribution and many data analysis techniques assume an approximate normal distribution. The third example depicts a *bimodal* distribution where the values cluster in two locations, in this case primarily at both ends of the distribution. The final three histograms show frequency distributions that either increase or decrease linearly as the values increase (fourth and fifth histogram) or have a nonlinear distribution as in the case of the sixth histogram where the number of observations is increasing exponentially as the values increase.

A frequency histogram can also tell us if there is something unusual about the variables. In Figure 2.7, the first histogram appears to contain two approximately normal distributions and leads us to question whether the data table contains two distinct types of observations, each with a separate

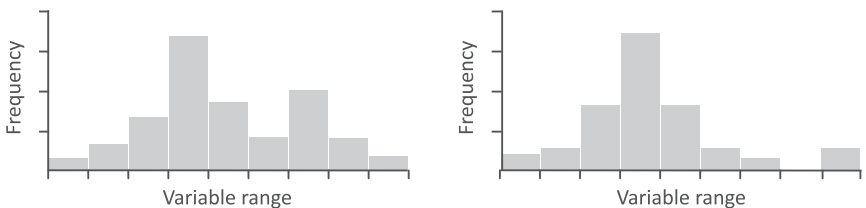


FIGURE 2.7 More complex frequency distributions.

frequency distribution. In the second histogram, there appears to be a small number of high values that do not follow the bell-shaped distribution that the majority of observations follow. In this case, it is possible that these values are errors and need to be further investigated.

2.5.3 Range

The range is a simple measure of the variation for a particular variable. It is calculated as the difference between the highest and lowest values. The following variable will be used to illustrate:

2, 3, 4, 6, 7, 7, 8, 9

The range is 7 calculated from the highest value (9) minus the lowest value (2). Ranges can be used with variables measured on an ordinal, interval, or ratio scale.

2.5.4 Quartiles

Quartiles divide a continuous variable into four even segments based on the number of observations. The first quartile (Q1) is at the 25% mark, the second quartile (Q2) is at the 50% mark, and the third quartile (Q3) is at the 75% mark. The calculation for Q2 is the same as the median value (described earlier). The following list of values is used to illustrate how quartiles are calculated:

3, 4, 7, 2, 3, 7, 4, 2, 4, 7, 4

The values are initially sorted:

2, 2, 3, 3, 4, 4, 4, 4, 7, 7, 7

Next, the median or Q2 is located in the center:

2, 2, 3, 3, 4, 4, 4, 4, 7, 7, 7

We now look for the center of the first half (shown underlined) or Q1:

2, 2, 3, 3, 4, 4, 4, 4, 7, 7, 7

The value of Q1 is recorded as 3.

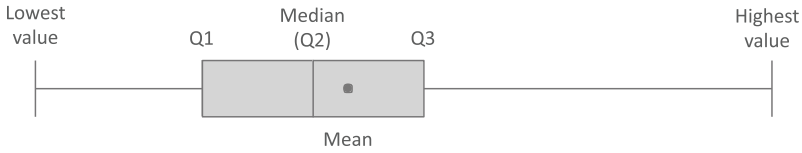


FIGURE 2.8 Overview of elements of a box plot.

Finally, we look for the center of the second half (shown underlined) or Q3:

2, 2, 3, 3, 4, 4, 4, 4, 7, 7, 7

The value of Q3 is identified as 7.

When the boundaries of the quartiles do not fall on a specific value, the quartile value is calculated based on the two numbers adjacent to the boundary. The *interquartile range* is defined as the range from Q1 to Q3. In this example it would be $7 - 3$ or 4.

2.5.5 Box Plots

Box plots provide a succinct summary of the overall frequency distribution of a variable. Six values are usually displayed: the lowest value, the lower quartile (Q1), the median (Q2), the upper quartile (Q3), the highest value, and the mean. In the conventional box plot displayed in Figure 2.8, the box in the middle of the plot represents where the central 50% of observations lie. A vertical line shows the location of the median value and a dot represents the location of the mean value. The horizontal line with a vertical stroke between “lowest value” and “Q1” and “Q3” and “highest value” are the “tails”—the values in the first and fourth quartiles.

Figure 2.9 provides an example of a box plot for one variable (*MPG*). The plot visually displays the lower (9) and upper (46.6) bounds of the variable. Fifty percent of observations begin at the lower quartile (17.5)

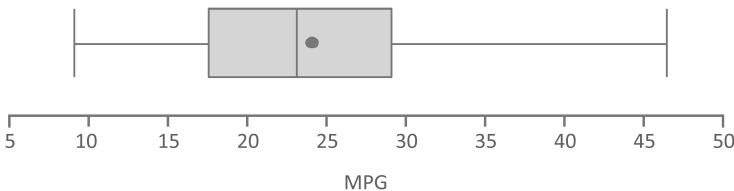


FIGURE 2.9 Box plot for the variable *MPG*.

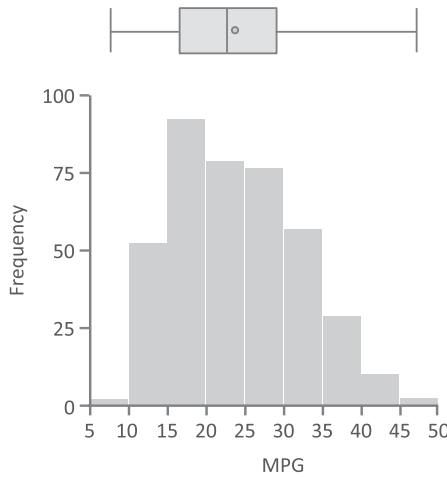


FIGURE 2.10 Comparison of frequency histogram and a box plot for the variable *MPG*.

and end at the upper quartile (29). The median and the mean values are close, with the mean slightly higher (around 23.6) than the median (23). Figure 2.10 shows a box plot and a histogram side-by-side to illustrate how the distribution of a variable is summarized using the box plot.

“Outliers,” the solitary data values close to the ends of the range of values, are treated differently in various forms of the box plot. Some box plots do not graphically separate them from the first and fourth quartile depicted by the horizontal lines that are to the left and the right of the box. In other forms of box plots, these extreme values are replaced with the highest and lowest values not considered an outlier and the outliers are explicitly drawn (using small circles) outside the main plot as shown in Figure 2.11.

Box plots help in understanding the symmetry of a frequency distribution. If both the mean and median have approximately the same value,

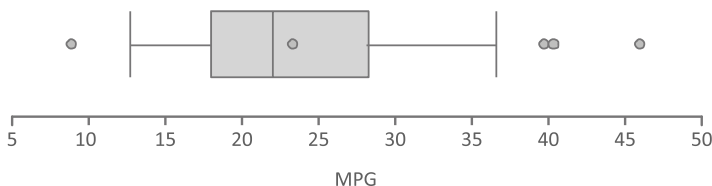


FIGURE 2.11 A box plot with extreme values explicitly shown as circles.

there will be about the same number of values above and below the mean and the distribution will be roughly symmetric.

2.5.6 Variance

The *variance* describes the spread of the data and measures how much the values of a variable differ from the mean. For variables that represent only a sample of some population and not the population as a whole, the variance formula is

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

The sample variance is referred to as s^2 . The actual value (x_i) minus the mean value (\bar{x}) is squared and summed for all values of a variable. This value is divided by the number of observations minus 1 ($n - 1$).

The following example illustrates the calculation of a variance for a particular variable:

$$3, 4, 4, 5, 5, 5, 6, 6, 6, 7, 7, 8, 9$$

where the mean is

$$\bar{x} = \frac{3 + 4 + 4 + 5 + 5 + 5 + 6 + 6 + 6 + 7 + 7 + 8 + 9}{13}$$

$$\bar{x} = 5.8$$

Table 2.3 is used to calculate the sum, using the mean value of 5.8.

To calculate s^2 , we substitute the values from Table 2.3 into the variance formula:

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

$$s^2 = \frac{34.32}{13 - 1}$$

$$s^2 = 2.86$$

The variance reflects the average squared deviation and can be calculated for variables measured on the interval or ratio scale.

TABLE 2.3 Variance Intermediate Steps

x	\bar{x}	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
3	5.8	-2.8	7.84
4	5.8	-1.8	3.24
4	5.8	-1.8	3.24
5	5.8	-0.8	0.64
5	5.8	-0.8	0.64
5	5.8	-0.8	0.64
6	5.8	0.2	0.04
6	5.8	0.2	0.04
6	5.8	0.2	0.04
7	5.8	1.2	1.44
7	5.8	1.2	1.44
8	5.8	2.2	4.84
9	5.8	3.2	10.24
			Sum = 34.32

2.5.7 Standard Deviation

The *standard deviation* is the square root of the variance. For a sample from a population, the formula is

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

where s is the sample standard deviation, x_i is the actual data value, \bar{x} is the mean for the variable, and n is the number of observations. For a calculated variance (e.g., 2.86) the standard deviation is calculated as $\sqrt{2.86}$ or 1.69.

The standard deviation is the most widely used measure of the deviation of a variable. The higher the value, the more widely distributed the variable's data values are around the mean. Assuming the frequency distribution is approximately normal (i.e., a bell-shaped curve), about 68% of all observations will fall within one standard deviation of the mean (34% less than and 34% greater than). For example, a variable has a mean value of 45 with a standard deviation value of 6. Approximately 68% of the observations should be in the range 39–51 ($45 \pm$ one standard deviation) and approximately 95% of all observations fall within two standard deviations

of the mean (between 33 and 57). Standard deviations can be calculated for variables measured on the interval or ratio scales.

It is possible to calculate a normalized value, called a *z-score*, for each data element that represents the number of standard deviations that element's value is from the mean. The following formula is used to calculate the *z-score*:

$$z = \frac{x_i - \bar{x}}{s}$$

where z is the *z-score*, x_i is the actual data value, \bar{x} is the mean for the variable, and s is the standard deviation. A *z-score* of 0 indicates that a data element's value is the same as the mean, data elements with *z-scores* greater than 0 have values greater than the mean, and elements with *z-scores* less than 0 have values less than the mean. The magnitude of the *z-score* reflects the number of standard deviations that value is from the mean. This calculation can be useful for comparing variables measured on different scales.

2.5.8 Shape

Previously in this chapter, we discussed ways to visualize the frequency distribution. In addition to these visualizations, there are methods for quantifying the lack of symmetry or *skewness* in the distribution of a variable. For asymmetric distributions, the bulk of the observations are either to the left or the right of the mean. For example, in Figure 2.12 the frequency distribution is asymmetric and more of the observations are to the left of the mean than to the right; the right tail is longer than the left tail. This is an example of a positive, or right skew. Similarly, a negative, or left skew would have more of the observations to the right of the mean value with a longer tail on the left.

It is possible to calculate a value for skewness that describes whether the variable is positively or negatively skewed and the degree of skewness. One formula for estimating skewness, where the variable is x with individual values x_i , and n data values is

$$skewness = \left(\frac{\sqrt{n \times (n - 1)}}{n - 2} \right) \times \frac{1/n \times \sum_{i=1}^n (x_i - \bar{x})^3}{\left(1/n \times \sum_{i=1}^n (x_i - \bar{x})^2 \right)^{3/2}}$$

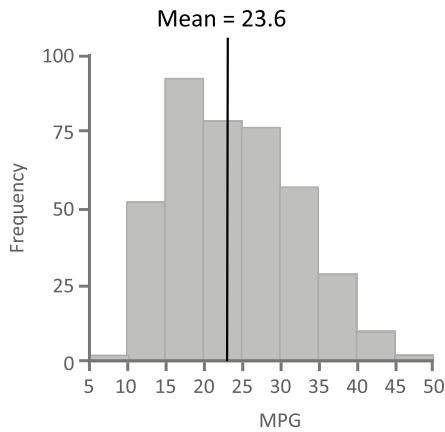


FIGURE 2.12 Frequency distribution showing a positive skew.

A skewness value of zero indicates a symmetric distribution. If the lower tail is longer than the upper tail the value is positive; if the upper tail is longer than the lower tail, the skewness score is negative. Figure 2.13 shows examples of skewness values for two variables. The variable *alkphos* in the plot on the left has a positive skewness value of 0.763, indicating that the majority of observations are to the left of the mean, whereas the negative skewness value for the variable *mcv* in the plot on the right indicates that the majority are to the right of the mean. That the skewness value for *mcv* is closer to zero than *alkphos* indicates that *mcv* is more symmetric than *alkphos*.

In addition to the symmetry of the distribution, the type of peak the distribution has should be considered and it can be characterized by a measurement called *kurtosis*. The following formula can be used for

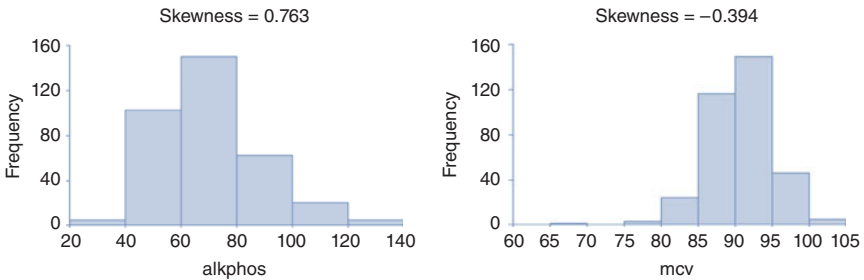


FIGURE 2.13 Skewness estimates for two variables.

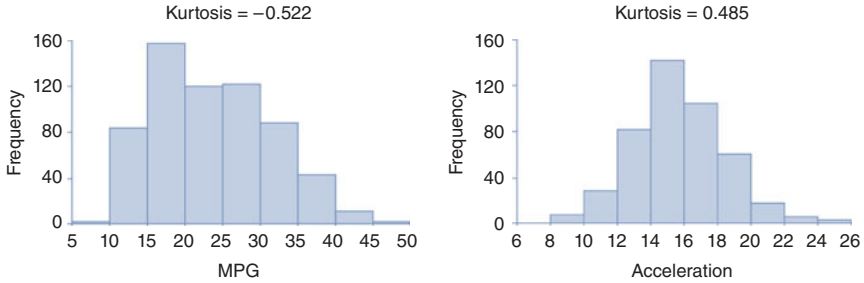


FIGURE 2.14 Kurtosis estimates for two variables.

calculating kurtosis for a variable x , where x_i represents the individual values, and n the number of data values:

$$kurtosis = \frac{n-1}{(n-2) \times (n-3)} \times \left((n+1) \times \frac{\sum_{i=1}^n (x_i - \bar{x})^4 / n}{\left(\sum_{i=1}^n (x_i - \bar{x})^2 / n \right)^2} - 3 \right) + 6$$

Variables with a pronounced peak near the mean have a high kurtosis score while variables with a flat peak have a low kurtosis score. Figure 2.14 illustrates kurtosis scores for two variables.

It is important to understand whether a variable has a normal distribution, since a number of data analysis approaches require variables to have this type of frequency distribution. Values for skewness and kurtosis close to zero indicate that the shape of a frequency distribution for a variable approximates a normal distribution which is important for checking assumptions in certain data analysis methods.

2.6 CONFIDENCE INTERVALS

Up to this point, we have been looking at ways of summarizing information on a set of randomly collected observations. This summary information is usually referred to as *statistics* as they summarize only a collection of observations that is a subset of a larger population. However, information derived from a sample of observations can only be an approximation of the entire population. To make a definitive statement about an entire *population*, every member of that population would need to be measured. For example, if we wanted to say for certain that the average weight of men in the United States is 194.7 lb, we would have to collect the weight measurements

for every man living in the United States and derive a mean from these observations. This is not possible or practical in most situations.

It is possible, however, to make estimates about a population by using *confidence intervals*. Confidence intervals are a measure of our uncertainty about the statistics we calculate from a single sample of observations. For example, the confidence interval might state that the average weight of men in the United States is between 191.2 lb and 198.2 lb to take into account the uncertainty of measuring only a sample of the total population. Only if the sample of observations is a truly random sample of the entire population can these types of estimates be made.

To understand how a statistic, such as the mean or mode, calculated from a single sample can reliably be used to infer a corresponding value of the population that cannot be measured and is therefore unknown, you need to understand something about *sample distributions*. Each statistic has a corresponding unknown value in the population called a *parameter* that can be estimated. In the example used in this section, we chose to calculate the statistic *mean* for the weight of US males. The mean value for the random sample selected is calculated to be 194.7 lb. If another random sample with the same number of observations were collected, the mean could also be calculated and it is likely that the means of the two samples would be close but not identical. If we take many random samples of equal size and calculate the mean value from each sample, we would begin to form a frequency distribution. If we were to take infinitely many samples of equal size and plot on a graph the value of the mean calculated from each sample, it would produce a *normal frequency distribution* that reflects the distribution of the sample means for the population mean under consideration. The distribution of a statistic computed for each of many random samples is called a *sampling distribution*. In our example, we would call this the *sampling distribution of the mean*.

Just as the distributions for a statistical variable discussed in earlier sections have a mean and a standard deviation, so also does the sampling distribution. However, to make clear when these measures are being used to describe the distribution of a statistic rather than the distribution of a variable, distinct names are used. The mean of a sampling distribution is called the *expected value of the mean*: it is the mean expected of the population. The standard deviation of the sampling distribution is called the *standard error*: it measures how much error to expect from equally sized random samples drawn from the same population. The standard error informs us of the average difference between the mean of a sample and the expected value.

The sample size is important. It is beyond the scope of this book to explain the details, but regardless of how the values of a variable for the population are distributed, the sampling distribution of a statistic calculated on samples from that variable will have a *normal* form when the size chosen for the samples has at least 30 observations. This is known as the *law of large numbers*, or more formally as the *central limit theorem*.

The standard error plays a fundamental role in inferential statistics by providing a measurable level of confidence in how well a sample mean estimates the mean of the population. The standard error can be calculated from a sample using the following formula:

$$\text{standard error of the sampling distribution} = \frac{s}{\sqrt{n}}$$

where s is the standard deviation of a sample and n is the number of observations in the sample. Because the size n is in the denominator and the standard deviation s is in the numerator, small samples with large variations increase the standard error, reducing the confidence that the sample statistic is a close approximation of the population parameter we are trying to estimate.

The data analyst or the team calculating the confidence interval should decide what the desired level of confidence should be. Confidence intervals are often based on a 95% confidence level, but sometimes a more stringent 99% confidence level or less stringent 90% level is used. Using a confidence interval of 95% to illustrate, one way to interpret this confidence level is that, on average, the correct population value will be found within the derived confidence interval 95 times out of every 100 samples collected. In these 100 samples, there will be 5 occasions on average when this value does not fall within the range. The confidence level is usually stated in terms of α from the following equation:

$$\text{confidence interval} = 100 \times (1 - \alpha)$$

For a 90% confidence level α is 0.1; for a 95% confidence level α is 0.05; for a 99% confidence level α is 0.01; and so on. The value used for this level of confidence will affect the size of the interval; that is, the higher the desired level of confidence the wider the confidence interval.

Along with the value of α selected, the confidence interval is based on the standard error. The estimated range or confidence interval is calculated using this confidence level along with information on the number of

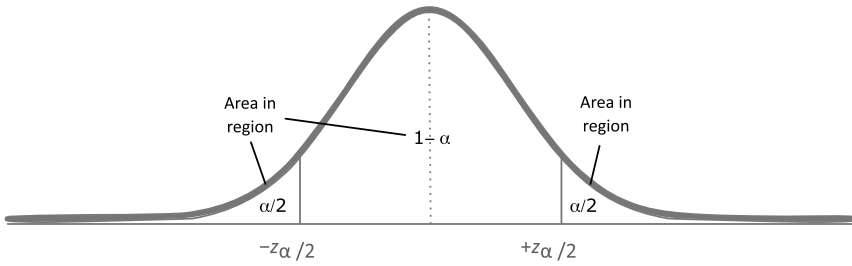


FIGURE 2.15 Illustration of the standard z -distribution to calculate $z_{\alpha/2}$.

observations in the sample as well as the variation in the sample's data. The formula showing a confidence interval for a mean value is shown here:

$$\bar{x} \pm z_{\alpha/2} \left(\frac{s}{\sqrt{n}} \right)$$

where \bar{x} is the mean value, s is the standard deviation, and n is the number of observations in the sample. The value for $z_{\alpha/2}$ is based on the area to the right of a standard z -distribution as illustrated in Figure 2.15 since the total area under this curve is 1. This number can be derived from a standard statistical table or computer program.

To illustrate, the fuel efficiency of 100 specific cars is measured and a mean value of 30.35 *MPG* is calculated with a standard deviation of 2.01. Using an alpha value of 0.05 (which translates into a $z_{\alpha/2}$ value of 1.96), the confidence interval is calculated as

$$\begin{aligned} & \bar{x} \pm z_{\alpha/2} \left(\frac{s}{\sqrt{n}} \right) \\ & 30.35 \pm 1.96 \left(\frac{2.01}{\sqrt{100}} \right) \end{aligned}$$

Hence, the confidence interval for the average fuel efficiency is 30.35 ± 0.393 or between 29.957 and 30.743.

For calculation of a confidence interval where sigma is unknown and the number of observations is less than 30 observations, a t -distribution should be used (see Urdan (2010), Anderson et al. (2010), Witte & Witte (2009), Kachigan (1991), Freedman et al. (2007), and Vickers (2010) for more details).

2.7 HYPOTHESIS TESTS

Hypothesis tests are used to support making decisions by helping to understand whether data collected from a sample of all possible observations supports a particular hypothesis. For example, a company manufacturing hair care products wishes to say that the average amount of shampoo within the bottles is 200 mL. To test this hypothesis, the company collects a random sample of 100 shampoo bottles and precisely measures the contents of the bottle. If it is inferred from the sample that the average amount of shampoo in each bottle is not 200 mL then a decision may be made to stop production and rectify the manufacturing problem.

The first step is to formulate the hypothesis that will be tested. This hypothesis is referred to as the null hypothesis (H_0). The null hypothesis is stated in terms of what would be expected if there were nothing unusual about the measured values of the observations in the data from the samples we collect—“null” implies the absence of effect. In the example above, if we expected each bottle of shampoo to contain 200 mL of shampoo, the null hypothesis would be: the average volume of shampoo in a bottle is 200 mL. Its corresponding alternative hypothesis (H_a) is that they differ or, stated in a way that can be measured, that the average is not equal to 200 mL. For this example, the null hypothesis and alternative hypothesis would be shown as

$$H_0 : \mu = 200$$

$$H_a : \mu \neq 200$$

This hypothesis will be tested using the sample data collected to determine whether the mean value is different enough to warrant rejecting the null hypothesis. Hence, the result of a hypothesis test is either to *fail to reject* or *reject* the null hypothesis. Since we are only looking at a sample of the observations—we are not testing every bottle being manufactured—it is impossible to make a statement about the average with total certainty. Consequently, it is possible to reach an incorrect conclusion. There are two categories of errors that can be made. One is to reject the null hypothesis when, in fact, the null hypothesis should stand (referred to as a type I error); the other is to accept the null hypothesis when it should be rejected (or type II error). The threshold of probability used to determine a type I error should be decided before performing the test. This threshold, which is also referred to as the level of significance or α , is often set to 0.05 (5% chance of a type I error); however, more stringent (such as 0.01 or 1%

chance) or less stringent values (such as 0.1 or 10% chance) can be used depending on the consequences of an incorrect decision.

The next step is to specify the standardized test statistic (T). We are interested in determining whether the average of the sample data we collected is either meaningfully or trivially different from the population average. Is it likely that we would find as great a difference from the population average were we to collect other random samples of the same size and compare their mean values? Because the hypothesis involves the mean, we use the following formula to calculate the test statistic:

$$T = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

where \bar{x} is the calculated mean value of the sample, μ_0 is the population mean that is the subject of the hypothesis test, s is the standard deviation of the sample, and n is the number of observations. (Recall that the denominator is the standard error of the sampling distribution.) In this example, the average shampoo bottle volume measured over the 100 samples (n) is 199.94 (\bar{x}) and the standard deviation is 0.613 (s).

$$T = \frac{199.94 - 200}{0.613 / \sqrt{100}} = -0.979$$

Assuming we are using a value for α of 0.05 as the significance level to formulate a decision rule to either let the null hypothesis stand or reject it, it is necessary to identify a range of values where 95% of all the sample means would lie. As discussed in Section 2.6, the law of large numbers applies to the sampling distribution of the statistic T : when there are at least 30 observations, the frequency distribution of the sample means is approximately normal and we can use this distribution to estimate regions to accept the null hypothesis. This region has two critical upper and lower bound values $C1$ and $C2$. Ninety-five percent of all sample means lie between these values and 5% lie outside these values (0.025 below $C2$ and 0.025 above $C1$) (see Figure 2.16). We reject the null hypothesis if the value of T is outside this range (i.e., greater than $C1$ or less than $C2$) or let the null hypothesis stand if it is inside the range.

Values for $C1$ and $C2$ can be calculated using a standard z -distribution table lookup and would be $C2 = -1.96$ and $C1 = +1.96$. These z -values were selected where the combined area to the left of $C2$ and to the right of $C1$ would equal 0.05.

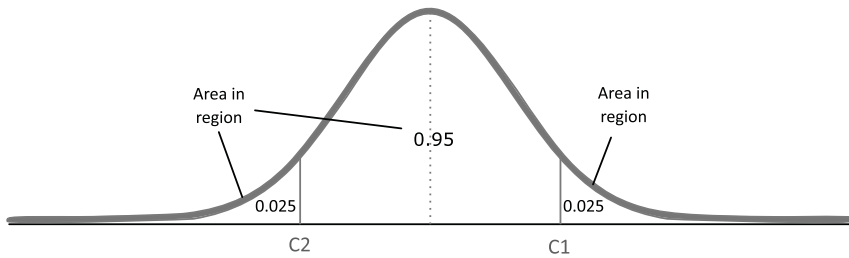


FIGURE 2.16 Standard z -distribution.

Since T is -0.979 and is greater than -1.96 and less than $+1.96$, we let the null hypothesis stand and conclude that the value is within the acceptable range.

In this example, the hypothesis test is referred to as a two-tailed test—that is, we tested the hypothesis for values above and below the critical values; however, hypothesis tests can be structured such that they are testing for values only above or below a value.

It is a standard practice to also calculate a p -value which corresponds to the probability of obtaining a test statistic value at least as extreme as the observed value (assuming the null hypothesis is true). This p -value can also be used to assess the null hypothesis, where the null hypothesis is rejected if it is less than the value of alpha. This value can be looked up using a standard z -distribution table as found by an online search or readily available software. In this example, the p -value would be 0.33 . Since this value is not less than 0.05 (as defined earlier) we again do not reject the null hypothesis.

EXERCISES

A set of 10 hypothetical patient records from a large database is presented in Table 2.4. Patients with a diabetes value of 1 have type-II diabetes and patients with a diabetes value of 0 do not have type-II diabetes.

1. For each of the following variables, assign them to one of the following scales: nominal, ordinal, interval, or ratio:
 - (a) *Name*
 - (b) *Age*
 - (c) *Gender*
 - (d) *Blood group*
 - (e) *Weight (kg)*

TABLE 2.4 Table of Patient Records

Name	Age	Gender	Blood Group	Weight (kg)	Height (m)	Systolic Blood Pressure (mmHg)	Diastolic Blood Pressure (mmHg)	Diabetes
P. Lee	35	Female	A Rh+	50	1.52	68	112	0
R. Jones	52	Male	O Rh-	115	1.77	110	154	1
J. Smith	45	Male	O Rh+	96	1.83	88	136	0
A. Patel	70	Female	O Rh-	41	1.55	76	125	0
M. Owen	24	Male	A Rh-	79	1.82	65	105	0
S. Green	43	Male	O Rh-	109	1.89	114	159	1
N. Cook	68	Male	A Rh+	73	1.76	108	136	0
W. Hands	77	Female	O Rh-	104	1.71	107	145	1
P. Rice	45	Female	O Rh+	64	1.74	101	132	0
F. Marsh	28	Male	O Rh+	136	1.78	121	165	1

- (f) Height (m)
- (g) Systolic blood pressure (mmHg)
- (h) Diastolic blood pressure (mmHg)
- (i) Diabetes

TABLE 2.5 Table with
Variables Name and Age

Name	Age
P. Lee	35
R. Jones	52
J. Smith	45
A. Patel	70
M. Owen	24
S. Green	43
N. Cook	68
W. Hands	77
P. Rice	45
F. Marsh	28

TABLE 2.6 Retail Transaction Data Set

Customer	Store	Product Category	Product Description	Sale Price (\$)	Profit (\$)
B. March	New York, NY	Laptop	DR2984	950	190
B. March	New York, NY	Printer	FW288	350	105
B. March	New York, NY	Scanner	BW9338	400	100
J. Bain	New York, NY	Scanner	BW9443	500	125
T. Goss	Washington, DC	Printer	FW199	200	60
T. Goss	Washington, DC	Scanner	BW39339	550	140
L. Nye	New York, NY	Desktop	LR21	600	60
L. Nye	New York, NY	Printer	FW299	300	90
S. Cann	Washington, DC	Desktop	LR21	600	60
E. Sims	Washington, DC	Laptop	DR2983	700	140
P. Judd	New York, NY	Desktop	LR22	700	70
P. Judd	New York, NY	Scanner	FJ3999	200	50
G. Hinton	Washington, DC	Laptop	DR2983	700	140
G. Hinton	Washington, DC	Desktop	LR21	600	60
G. Hinton	Washington, DC	Printer	FW288	350	105
G. Hinton	Washington, DC	Scanner	BW9443	500	125
H. Fu	New York, NY	Desktop	ZX88	450	45
H. Taylor	New York, NY	Scanner	BW9338	400	100

2. Calculate the following statistics for the variable *Age* (from Table 2.5):
 - (a) Mode
 - (b) Median
 - (c) Mean
 - (d) Range
 - (e) Variance
 - (f) Standard deviation
3. Using the data in Table 2.6, create a histogram of *Sale Price* (\$) using the following intervals: 0 to less than 250, 250 to less than 500, 500 to less than 750, and 750 to less than 1000.

FURTHER READING

A number of books provide basic introductions to statistical methods including Donnelly (2007) and Levine & Stephan (2010). Numerous books provide additional details on the descriptive and inferential statistics as, for example, Urdan (2010), Anderson et al. (2010), Witte & Witte (2009), Kachigan (1991), Freedman et al. (2007), and Vickers (2010). The conceptual difference between standard error and standard deviation described in Sections 2.6 and 2.7 is often difficult to grasp. For further discussion, see the section on sampling distributions in Kachigan (1991) and the chapter on standard error in Vickers (2010). For further reading on communicating information, see Tufte (1990, 1997a, 1997b, 2001, 2006). These works describe a theory of data graphics and information visualization that are illustrated by many examples.