CHAPTER 2

TESTING DATA FOR NORMALITY

2.1 OBJECTIVES

In this chapter, you will learn the following items:

- How to find a data sample's kurtosis and skewness and determine if the sample meets acceptable levels of normality.
- How to use SPSS® to find a data sample's kurtosis and skewness and determine if the sample meets acceptable levels of normality.
- How to perform a Kolmogorov–Smirnov one-sample test to determine if a data sample meets acceptable levels of normality.
- How to use SPSS to perform a Kolmogorov–Smirnov one-sample test to determine if a data sample meets acceptable levels of normality.

2.2 INTRODUCTION

Parametric statistical tests, such as the *t*-test and one-way analysis of variance, are based on particular assumptions or parameters. The data samples meeting those parameters are randomly drawn from a normal population, based on independent observations, measured with an interval or ratio scale, possess an adequate sample size (see Chapter 1), and approximately resemble a normal distribution. Moreover, comparisons of samples or variables should have approximately equal variances. If data samples violate one or more of these assumptions, you should consider using a nonparametric test.

Examining the data gathering method, scale type, and size of a sample are fairly straightforward. However, examining a data sample's resemblance to a normal distribution, or its normality, requires a more involved analysis. Visually inspecting a graphical representation of a sample, such as a stem and leaf plot or a box and whisker plot, might be the most simplistic examination of normality. Statisticians advocate this technique in beginning statistics; however, this measure of normality does not suffice for strict levels of defensible analyses.

Nonparametric Statistics: A Step-by-Step Approach, Second Edition. Gregory W. Corder and Dale I. Foreman.

In this chapter, we present three quantitative measures of sample normality. First, we discuss the properties of the normal distribution. Then, we describe how to examine a sample's kurtosis and skewness. Next, we describe how to perform and interpret a Kolmogorov–Smirnov one-sample test. In addition, we describe how to perform each of these procedures using SPSS.

2.3 DESCRIBING DATA AND THE NORMAL DISTRIBUTION

An entire chapter could easily be devoted to the description of data and the normal distribution and many books do so. However, we will attempt to summarize the concept and begin with a practical approach as it applies to data collection.

In research, we often identify some population we wish to study. Then, we strive to collect several independent, random measurements of a particular variable associated with our population. We call this set of measurements a *sample*. If we used good experimental technique and our sample adequately represents our population, we can study the sample to make inferences about our population. For example, during a routine checkup, your physician draws a sample of your blood instead of all of your blood. This blood sample allows your physician to evaluate all of your blood even though he or she only tested the sample. Therefore, all of your body's blood cells represent the population about which your physician makes an inference using only the sample.

While a blood sample leads to the collection of a very large number of blood cells, other fields of study are limited to small sample sizes. It is not uncommon to collect less than 30 measurements for some studies in the behavioral and social sciences. Moreover, the measurements lie on some scale over which the measurements vary about the mean value. This notion is called *variance*. For example, a researcher uses some instrument to measure the intelligence of 25 children in a math class. It is highly unlikely that every child will have the same intelligence level. In fact, a good instrument for measuring intelligence should be sensitive enough to measure differences in the levels of the children.

The variance s^2 can be expressed quantitatively. It can be calculated using Formula 2.1:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \tag{2.1}$$

where x_i is an individual value in the distribution, $\bar{x}$ is the distribution's mean, and n is the number of values in the distribution

As mentioned in Chapter 1, parametric tests assume that the variances of samples being compared are approximately the same. This idea is called homogeneity of variance. To compare sample variances, Field (2005) suggested that we obtain a variance ratio by taking the largest sample variance and dividing it by the smallest sample variance. The variance ratio should be less than 2. Similarly, Pett (1997) indicated that no sample's variance be twice as large as any other sample's variance. If the homogeneity of variance assumption cannot be met, one would use a nonparametric test.

A more common way of expressing a sample's variability is with its standard deviation, s. Standard deviation is the square root of variance where $s = \sqrt{s^2}$. In other words, standard deviation is calculated using Formula 2.2:

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} \tag{2.2}$$

As illustrated in Figure 2.1, a small standard deviation indicates that a sample's values are fairly concentrated about its mean, whereas a large standard deviation indicates that a sample's values are fairly spread out.

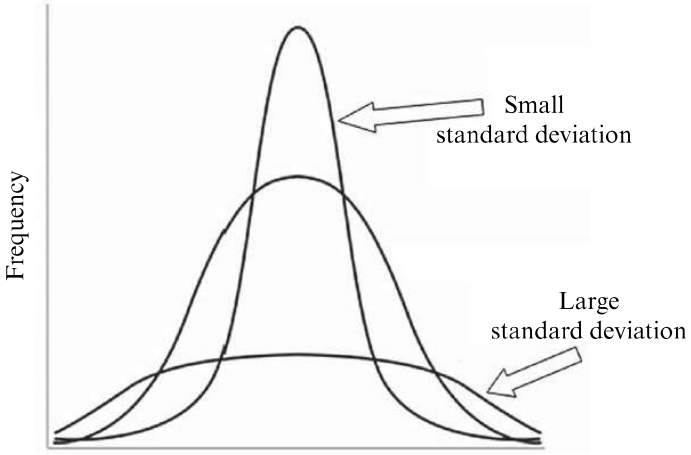

FIGURE 2.1

A histogram is a useful tool for graphically illustrating a sample's frequency distribution and variability (see Fig. 2.2). This graph plots the value of the measurements horizontally and the frequency of each particular value vertically. The middle value is called the median and the greatest frequency is called the mode.

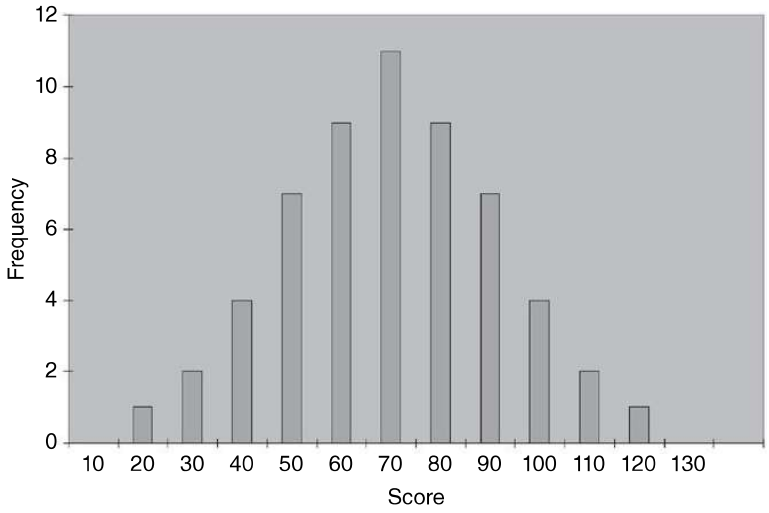

FIGURE 2.2

The mean and standard deviation of one distribution differ from the next. If we want to compare two or more samples, then we need some type of standard. A standard score is a way we can compare multiple distributions. The standard score that we use is called a z-score, and it can be calculated using Formula 2.3:

$$z = \frac{x_i - \overline{x}}{s} \tag{2.3}$$

where x_i is an individual value in the distribution, $\overline{x}$ is the distribution's mean, and s is the distribution's standard deviation.

There is a useful relationship between the standard deviation and z-score. We can think of the standard deviation as a unit of horizontal distance away from the mean on the histogram. One standard deviation from the mean is the same as $z = 1.0$. Two standard deviations from the mean are the same as $z = 2.0$. For example, if $s = 10$ and $\overline{x} = 70$ for a distribution, then $z = 1.0$ at $x = 80$ and $z = 2.0$ at $x = 90$. What is more, z-scores that lie below the mean have negative values. Using our example, $z = -1.0$ at $x = 60$ and $z = -2.0$ at $x = 50$. Moreover, $z = 0.0$ at the mean value, $x = 70$. These z-scores can be used to compare our distribution with another distribution, even if the mean and standard deviation are different. In other words, we can compare multiple distributions in terms of z-scores.

To this point, we have been focused on distributions with finite numbers of values, n. As more data values are collected for a given distribution, the histogram begins to resemble a bell shape called the normal curve. Figure 2.3 shows the relationship among the raw values, standard deviation, and z-scores of a population. Since we are describing a population, we use sigma, σ, to represent standard deviation and mu, μ, to represent the mean.

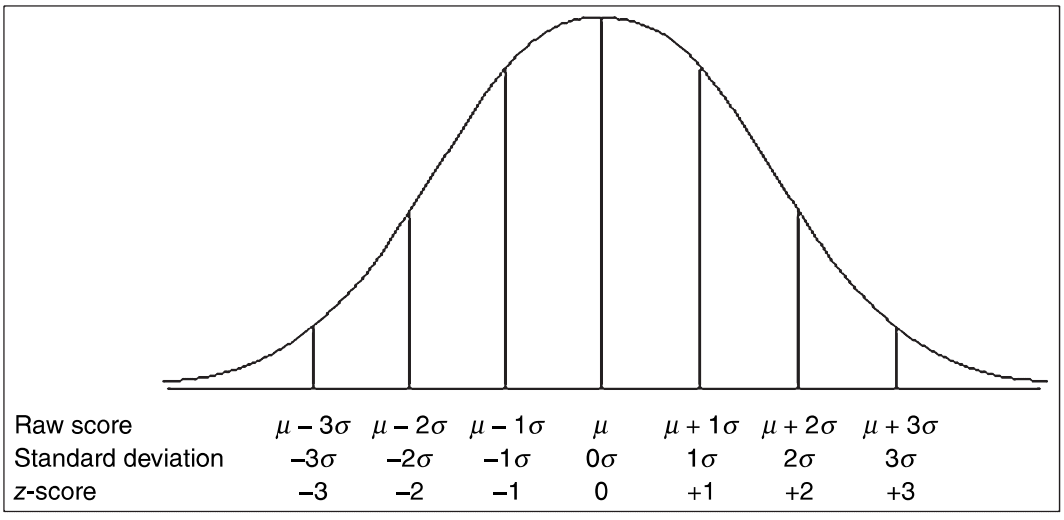

FIGURE 2.3

The normal curve has three particular properties (see Fig. 2.4). First, the mean, median, and mode are equal. Thus, most of the values lie in the center of the distribution. Second, the curve displays perfect symmetry about the mean. Third, the left and right sides of the curve, called the tails, are asymptotic. This means that they approach the horizontal axis, but never touch it.

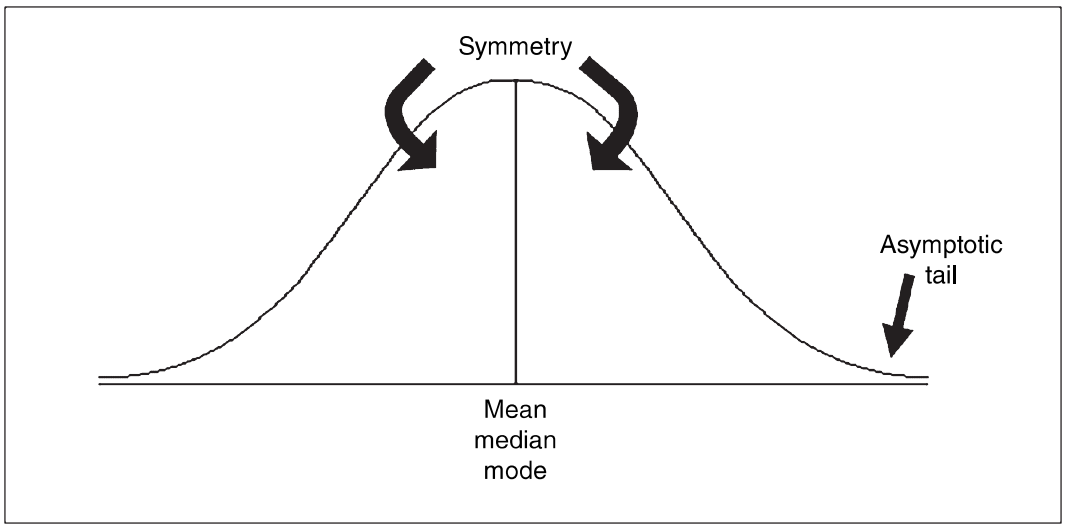

FIGURE 2.4

When we use a normal curve to represent probabilities p, we refer to it as the normal distribution. We set the area under the curve equal to $p = 1.0$. Since the distribution is symmetrical about the mean, $p = 0.50$ on the left side of the mean and $p = 0.50$ on the right. In addition, the ordinate of the normal curve, y, is the height of the curve at a particular point. The ordinate is tallest at the curve's center and decreases as you move away from the center. Table B.1 in Appendix B provides the z-scores, probabilities, and ordinates for the normal distribution.

2.4 COMPUTING AND TESTING KURTOSIS AND SKEWNESS FOR SAMPLE NORMALITY

A frequency distribution that resembles a normal curve is approximately normal. However, not all frequency distributions have the approximate shape of a normal curve. The values might be densely concentrated in the center or substantially spread out. The shape of the curve may lack symmetry with many values concentrated on one side of the distribution. We use the terms kurtosis and skewness to describe these conditions, respectively.

Kurtosis is a measure of a sample or population that identifies how flat or peaked it is with respect to a normal distribution. Stated another way, kurtosis refers to how concentrated the values are in the center of the distribution. As shown in Figure 2.5, a peaked distribution is said to be leptokurtic. A leptokurtic distribution has a positive kurtosis. If a distribution is flat, it is said to be platykurtic. A platykurtic distribution has a negative kurtosis.

The skewness of a sample can be described as a measure of horizontal symmetry with respect to a normal distribution. As shown in Figure 2.6, if a distribution's scores are concentrated on the right side of the curve, it is said to be left skewed. A left skewed distribution has a negative skewness. If a distribution's scores are concentrated on the left side of the curve, it is said to be right skewed. A right skewed distribution has a positive skewness.

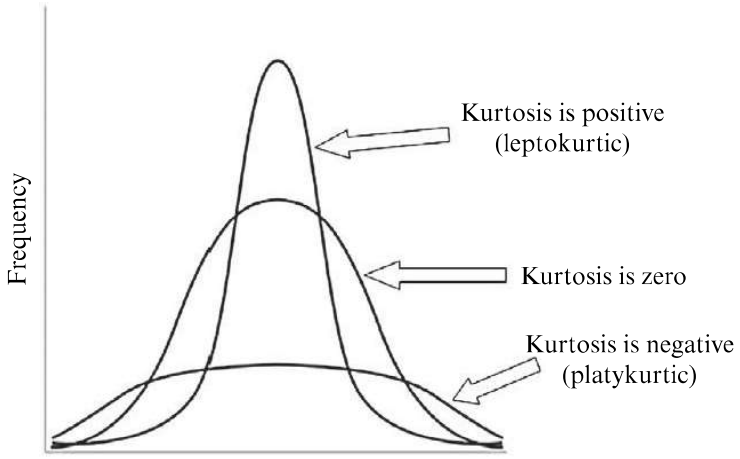

FIGURE 2.5

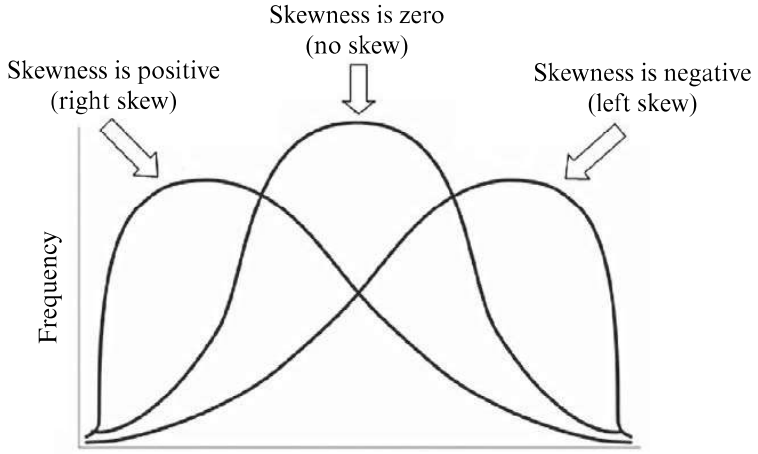

FIGURE 2.6

The kurtosis and skewness can be used to determine if a sample approximately resembles a normal distribution. There are five steps for examining sample normality in terms of kurtosis and skewness.

1. Determine the sample's mean and standard deviation.
2. Determine the sample's kurtosis and skewness.
3. Calculate the standard error of the kurtosis and the standard error of the skewness.
4. Calculate the z-score for the kurtosis and the z-score for the skewness.
5. Compare the z-scores with the critical region obtained from the normal distribution.

The calculations to find the values for a distribution's kurtosis and skewness require you to first find the sample mean $\bar{x}$ and the sample standard deviation s. Recall that standard deviation is found using Formula 2.2. The mean is found using Formula 2.4:

$$\bar{x} = \frac{\sum x_i}{n} \tag{2.4}$$

where $\sum x_i$ is the sum of the values in the sample and n is the number of values in the sample.

The kurtosis K and standard error of the kurtosis, SE_K, are found using Formula 2.5 and Formula 2.6:

$$K = \left[\frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum\left(\frac{x_i - \bar{x}}{s}\right)^4\right] - \frac{3(n-1)^2}{(n-2)(n-3)} \tag{2.5}$$

and

$$SE_K = \sqrt{\frac{24n(n-1)^2}{(n-2)(n-3)(n+5)(n+3)}} \tag{2.6}$$

The skewness S_k and standard error of the skewness, SE_{S_k}, are found using Formula 2.7 and Formula 2.8:

$$S_k = \frac{n}{(n-1)(n-2)}\sum\left(\frac{x_i - \bar{x}}{s}\right)^3 \tag{2.7}$$

$$SE_{S_k} = \sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}} \tag{2.8}$$

Normality can be evaluated using the z-score for the kurtosis, z_K, and the z-score for the skewness, z_{S_k}. Use Formula 2.9 and Formula 2.10 to find those z-scores:

$$z_K = \frac{K - 0}{SE_K} \tag{2.9}$$

$$z_{S_k} = \frac{S_k - 0}{SE_{S_k}} \tag{2.10}$$

Compare these z-scores with the values of the normal distribution (see Table B.1 in Appendix B) for a desired level of confidence α. For example, if you set $\alpha = 0.05$, then the calculated z-scores for an approximately normal distribution must fall between -1.96 and $+1.96$.

2.4.1 Sample Problem for Examining Kurtosis

The scores in Table 2.1 represent students' quiz performance during the first week of class. Use $\alpha = 0.05$ for your desired level of confidence. Determine if the samples of week 1 quiz scores are approximately normal in terms of its kurtosis.

TABLE 2.1

Week 1 quiz scores		
90	72	90
64	95	89
74	88	100
77	57	35
100	64	95
65	80	84
90	100	76

First, find the mean of the sample:

$$\overline{x} = \frac{\sum x_i}{n} = \frac{1706}{21}$$
$$\overline{x} = 80.24$$

Next, find the standard deviation. It is helpful to set up Table 2.2 to manage the summation when computing the standard deviation (see Formula 2.2):

$$s = \sqrt{\frac{\sum (x_i - \overline{x})^2}{n-1}} = \sqrt{\frac{5525.81}{21-1}} = \sqrt{276.29}$$
$$s = 16.62$$

Use the values for the mean and standard deviation to find the kurtosis. Again, it is helpful to set up Table 2.3 to manage the summation when computing the kurtosis (see Formula 2.5).

TABLE 2.2

x_i	$x_i - \overline{x}$	$(x_i - \overline{x})^2$
90	9.76	95.29
72	−8.24	67.87
90	9.76	95.29
64	−16.24	263.68
95	14.76	217.91
89	8.76	76.77
74	−6.24	38.91
88	7.76	60.25
100	19.76	390.53
77	−3.24	10.49
57	−23.24	540.01

TABLE 2.2 (*Continued*)

x_i	$x_i - \overline{x}$	$(x_i - \overline{x})^2$
35	−45.24	2046.49
100	19.76	390.53
64	−16.24	263.68
95	14.76	217.91
65	−15.24	232.20
80	−0.24	0.06
84	3.76	14.15
90	9.76	95.29
100	19.76	390.53
76	−4.24	17.96
		$\sum(x_i - \overline{x})^2 = 5525.81$

TABLE 2.3

x_i	$\frac{x_i - \overline{x}}{s}$	$\left(\frac{x_i - \overline{x}}{s}\right)^4$
90	0.587	0.119
72	−0.496	0.060
90	0.587	0.119
64	−0.977	0.911
95	0.888	0.622
89	0.527	0.077
74	−0.375	0.020
88	0.467	0.048
100	1.189	1.998
77	−0.195	0.001
57	−1.398	3.820
35	−2.722	54.864
100	1.189	1.998
64	−0.977	0.911
95	0.888	0.622
65	−0.917	0.706
80	−0.014	0.000
84	0.226	0.003
90	0.587	0.119
100	1.189	1.998
76	−0.255	0.004
		$\sum\left(\frac{x_i - \overline{x}}{s}\right)^4 = 69.020$

Compute the kurtosis:

$$
\begin{aligned}
K &= \left[\frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum\left(\frac{x_i-\bar{x}}{s}\right)^4\right]-\frac{3(n-1)^2}{(n-2)(n-3)} \\
&= \left[\frac{21(21+1)}{(21-1)(21-2)(21-3)}(69.020)\right]-\frac{3(21-1)^2}{(21-2)(21-3)} \\
&= \left[\frac{21(22)}{(20)(19)(18)}(69.020)\right]-\frac{3(20)^2}{(19)(18)} \\
&= [0.0675(69.020)]-3.509 = 4.662-3.509 \\
K &= 1.153
\end{aligned}
$$

Next, find the standard error of the kurtosis:

$$
\begin{aligned}
SE_K &= \sqrt{\frac{24n(n-1)^2}{(n-2)(n-3)(n+5)(n+3)}} \\
&= \sqrt{\frac{24(21)(21-1)^2}{(21-2)(21-3)(21+5)(21+3)}} \\
&= \sqrt{\frac{24(21)(20)^2}{(19)(18)(26)(24)}} = \sqrt{\frac{201{,}600}{213{,}408}} = \sqrt{0.945} \\
SE_K &= 0.972
\end{aligned}
$$

Finally, use the kurtosis and the standard error of the kurtosis to find a z-score:

$$
\begin{aligned}
z_K &= \frac{K-0}{SE_K} = \frac{1.153-0}{0.972} \\
z_K &= 1.186
\end{aligned}
$$

Use the z-score to examine the sample's approximation to a normal distribution. This value must fall between -1.96 and $+1.96$ to pass the normality assumption for $\alpha = 0.05$. Since this z-score value does fall within that range, the sample has passed our normality assumption for kurtosis. Next, the sample's skewness must be checked for normality.

2.4.2 Sample Problem for Examining Skewness

Based on the same values from the example listed earlier, determine if the samples of week 1 quiz scores are approximately normal in terms of its skewness.

Use the mean and standard deviation from the previous example to find the skewness. Set up Table 2.4 to manage the summation in the skewness formula.

Compute the skewness:

TABLE 2.4

x_i	$\frac{x_i-\bar{x}}{s}$	$\left(\frac{x_i-\bar{x}}{s}\right)^3$
90	0.587	0.203
72	−0.496	−0.122
90	0.587	0.203
64	−0.977	−0.932
95	0.888	0.700
89	0.527	0.146
74	−0.375	−0.053
88	0.467	0.102
100	1.189	1.680
77	−0.195	−0.007
57	−1.398	−2.732
35	−2.722	−20.159
100	1.189	1.680
64	−0.977	−0.932
95	0.888	0.700
65	−0.917	−0.770
80	−0.014	0.000
84	0.226	0.012
90	0.587	0.203
100	1.189	1.680
76	−0.255	−0.017
		$\sum\left(\frac{x_i-\bar{x}}{s}\right)^3 = -18.415$

$$S_k = \frac{n}{(n-1)(n-2)}\sum\left(\frac{x_i-\bar{x}}{s}\right)^3 = \frac{21}{(21-1)(21-2)}(-18.415)$$
$$= \frac{21}{(20)(19)}(-18.415)$$
$$S_k = -1.018$$

Next, find the standard error of the skewness:

$$SE_{S_k} = \sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}} = \sqrt{\frac{6(21)(21-1)}{(21-2)(21+1)(21+3)}}$$
$$= \sqrt{\frac{6(21)(20)}{(19)(22)(24)}} = \sqrt{\frac{2520}{10{,}032}} = \sqrt{0.251}$$
$$SE_{S_k} = 0.501$$

Finally, use the skewness and the standard error of the skewness to find a z-score:

$$z_{S_k} = \frac{S_k - 0}{SE_{S_k}} = \frac{-1.018}{0.501}$$

$$z_{S_k} = -2.032$$

Use the z-score to examine the sample's approximation to a normal distribution. This value must fall between -1.96 and $+1.96$ to pass the normality assumption for $\alpha = 0.05$. Since this z-score value does not fall within that range, the sample has failed our normality assumption for skewness. Therefore, either the sample must be modified and rechecked or you must use a nonparametric statistical test.

2.4.3 Examining Skewness and Kurtosis for Normality Using SPSS

We will analyze the examples earlier using SPSS.

2.4.3.1 Define Your Variables First, click the "Variable View" tab at the bottom of your screen. Then, type the name of your variable(s) in the "Name" column. As shown in Figure 2.7, we have named our variable "Wk1_Qz."

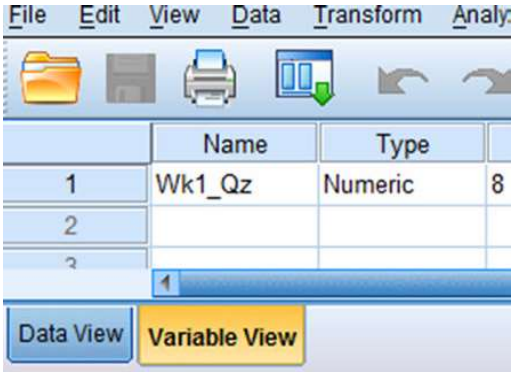

FIGURE 2.7

2.4.3.2 Type in Your Values Click the "Data View" tab at the bottom of your screen and type your data under the variable names. As shown in Figure 2.8, we have typed the values for the "Wk1_Qz" sample.

2.4.3.3 Analyze Your Data As shown in Figure 2.9, use the pull-down menus to choose "Analyze," "Descriptive Statistics," and "Descriptives . . ."

Choose the variable(s) that you want to examine. Then, click the button in the middle to move the variable to the "Variable(s)" box, as shown in Figure 2.10. Next, click the "Options . . ." button to open the "Descriptives: Options" window shown in Figure 2.11. In the "Distribution" section, check the boxes next to "Kurtosis" and "Skewness." Then, click "Continue."

Finally, once you have returned to the "Descriptives" window, as shown in Figure 2.12, click "OK" to perform the analysis.

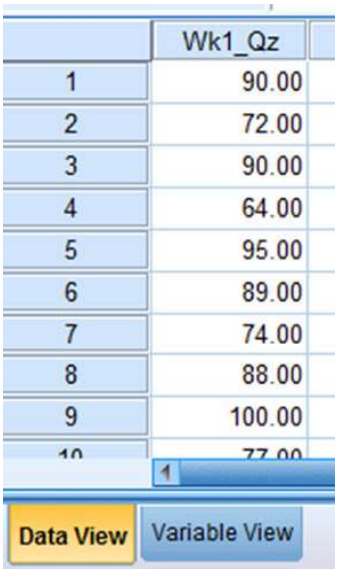

FIGURE 2.8

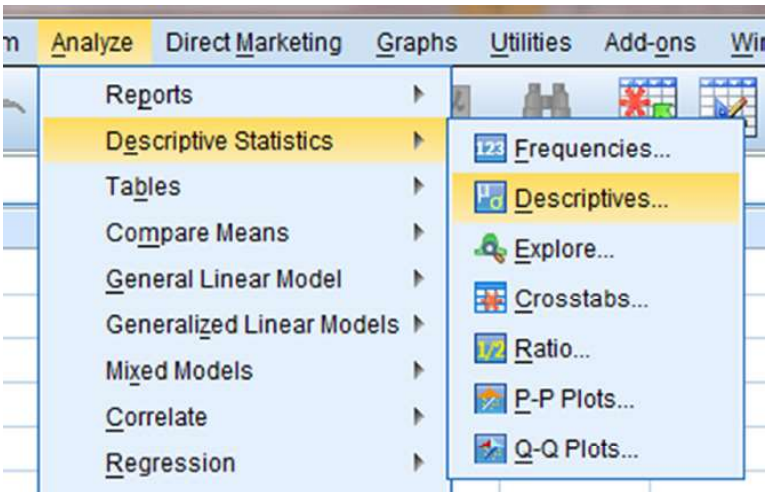

FIGURE 2.9

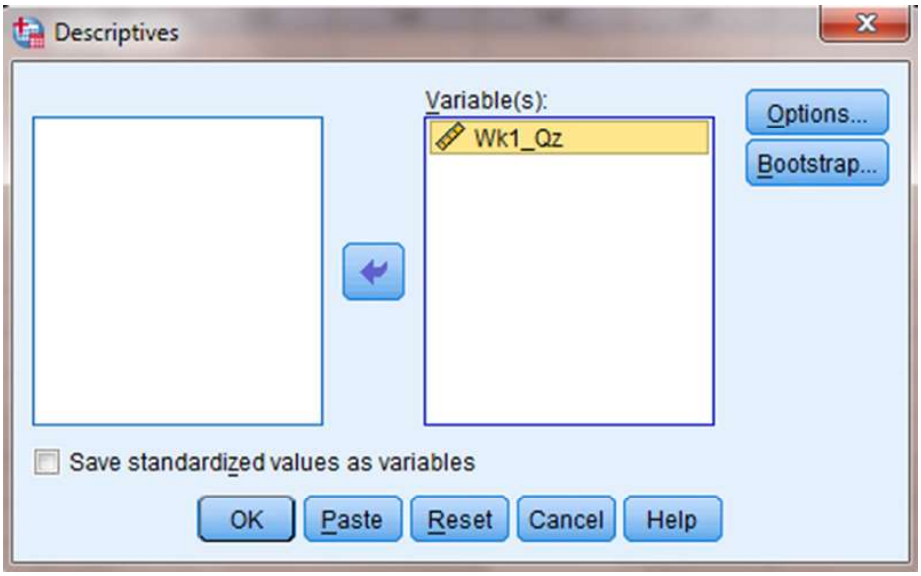

FIGURE 2.10

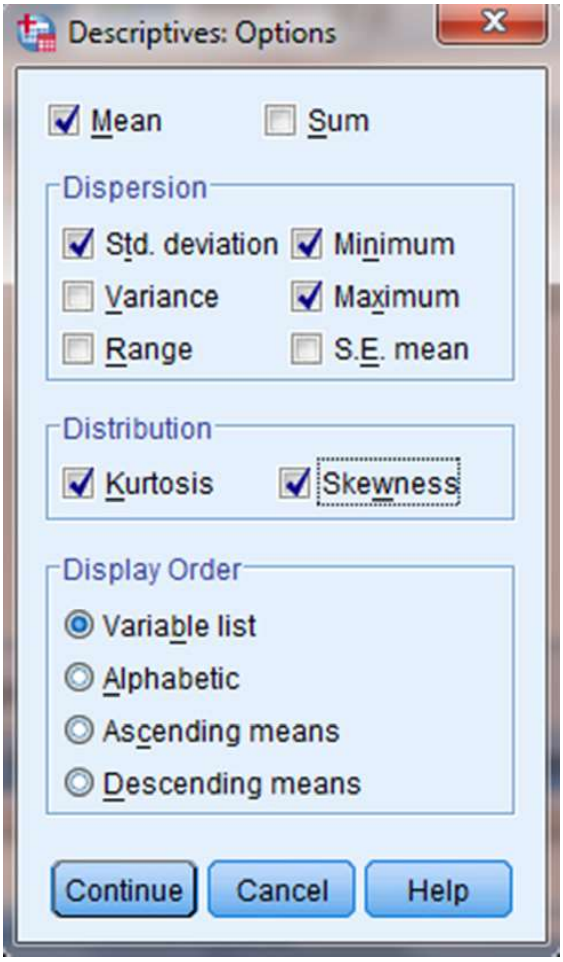

FIGURE 2.11

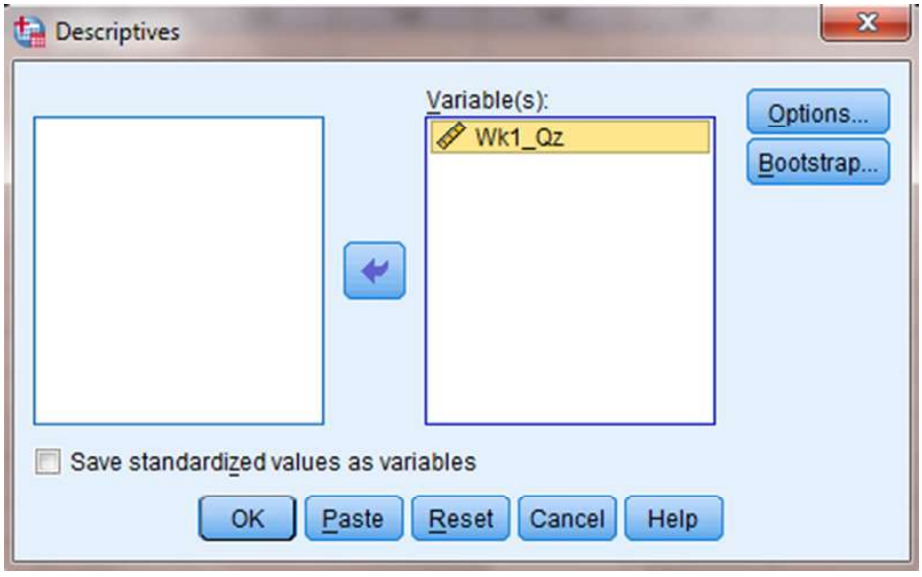

FIGURE 2.12

2.4.3.4 Interpret the Results from the SPSS Output Window The SPSS Output 2.1 provides the kurtosis and the skewness, along with their associated standard errors. In our example, the skewness is -1.018 and its standard error is 0.501. The kurtosis is 1.153 and its standard error is 0.972.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Wk1_Qz	21	35.00	100.00	80.2381	16.62199	-1.018	.501	1.153	.972
Valid N (listwise)	21								

SPSS OUTPUT 2.1

At this stage, we need to manually compute the z-scores for the skewness and kurtosis as we did in the previous examples. First, compute the z-score for kurtosis:

$$z_K = \frac{K-0}{SE_K} = \frac{1.153-0}{0.972}$$

$$z_K = 1.186$$

Next, we compute the z-score for skewness:

$$z_{S_k} = \frac{S_k-0}{SE_{S_k}} = \frac{-1.018}{0.501}$$

$$z_{S_k} = -2.032$$

Both of these values must fall between -1.96 and $+1.96$ to pass the normality assumption for $\alpha = 0.05$. The z-score for kurtosis falls within the desired range, but the z-score for skewness does not. Using $\alpha = 0.05$, the sample has passed the normality assumption for kurtosis, yet failed the normality assumption for skewness. Therefore, either the sample must be modified and rechecked or you must use a nonparametric statistical test.

2.5 COMPUTING THE KOLMOGOROV–SMIRNOV ONE-SAMPLE TEST

The Kolmogorov–Smirnov one-sample test is a procedure to examine the agreement between two sets of values. For our purposes, the two sets of values compared are an observed frequency distribution based on a randomly collected sample and an empirical frequency distribution based on the sample's population. Furthermore, the observed sample is examined for normality when the empirical frequency distribution is based on a normal distribution.

The Kolmogorov–Smirnov one-sample test compares two cumulative frequency distributions. A cumulative frequency distribution is useful for finding the number of observations above or below a particular value in a data sample. It is calculated by taking a given frequency and adding all the preceding frequencies

in the list. In other words, it is like making a running total of the frequencies in a distribution. Creating cumulative frequency distributions of the observed and empirical frequency distributions allow us to find the point at which these two distributions show the largest divergence. Then, the test uses the largest divergence to identify a two-tailed probability estimate p to determine if the samples are statistically similar or different.

To perform the Kolmogorov–Smirnov one-sample test, we begin by determining the relative empirical frequency distribution $\hat{f}_{x_i}$ based on the observed sample. This relative empirical frequency distribution will approximate a normal distribution since we are examining our observed values for sample normality. First, calculate the observed frequency distribution's midpoint M and standard deviation s. The midpoint and standard deviation are found using Formula 2.11 and Formula 2.12:

$$M = (x_{\max} + x_{\min}) \div 2 \tag{2.11}$$

where $x_{\max}$ is the largest value in the sample and $x_{\min}$ is the smallest value in the sample, and

$$s = \sqrt{\frac{\sum (f_i x_i^2) - \frac{\left(\sum f_i x_i\right)^2}{n}}{n-1}} \tag{2.12}$$

where x_i is a given value in the observed sample, f_i is the frequency of a given value in the observed sample, and n is the number of values in the observed sample.

Next, use the midpoint and standard deviation to calculate the z-scores (see Formula 2.13) for the sample values x_i,

$$z = \left| \frac{x_i - M}{s} \right| \tag{2.13}$$

Use those z-scores and Table B.1 in Appendix B to determine the probability associated with each sample value, $\hat{P}_{x_i}$. These p-values are the relative frequencies of the empirical frequency distribution $\hat{f}_r$.

Now, we find the relative values of the observed frequency distribution f_r. Use Formula 2.14:

$$f_r = \frac{f_i}{n} \tag{2.14}$$

where f_i is the frequency of a given value in the observed sample and n is the number of values in the observed sample.

Since the Kolmogorov–Smirnov test uses cumulative frequency distributions, both the relative empirical frequency distribution and relative observed frequency distribution must be converted into cumulative frequency distributions $\hat{F}_{x_i}$ and S_{x_i}, respectively. Use Formula 2.15 and Formula 2.16 to find the absolute value divergence $\tilde{D}$ and D between the cumulative frequency distributions:

$$\tilde{D} = \left| \hat{F}_{x_i} - S_{x_i} \right| \tag{2.15}$$

$$D = \left| \hat{F}_{x_i} - S_{x_{i-1}} \right| \tag{2.16}$$

Use the largest divergence with Formula 2.17 to calculate the Kolmogorov–Smirnov test statistic Z:

$$Z = \sqrt{n}\max\left(|D|, |\tilde{D}|\right) \tag{2.17}$$

Then, use the Kolmogorov–Smirnov test statistic Z and the Smirnov (1948) formula (see Formula 2.18, Formula 2.19, Formula 2.20, Formula 2.21, Formula 2.22, and Formula 2.23) to find the two-tailed probability estimate p:

$$\text{if } 0 \leq Z < 0.27, \text{ then } p = 1 \tag{2.18}$$

$$\text{if } 0.27 \leq Z < 1, \text{ then } p = 1 - \frac{2.506628}{Z}(Q + Q^9 + Q^{25}) \tag{2.19}$$

where

$$Q = e^{-1.233701Z^{-2}} \tag{2.20}$$

$$\text{if } 1 \leq Z < 3.1, \text{ then } p = 2(Q - Q^4 + Q^9 - Q^{16}) \tag{2.21}$$

where

$$Q = e^{-2Z^2} \tag{2.22}$$

$$\text{if } Z \geq 3.1, \text{ then } p = 0 \tag{2.23}$$

A p-value that exceeds the level of risk associated with the null hypothesis indicates that the observed sample approximates the empirical sample. Since our empirical distributions approximated a normal distribution, we can state that our observed sample is sufficiently normal for parametric statistics. Conversely, a p-value that is smaller than the level of risk indicates an observed sample that is not sufficiently normal for parametric statistics. The nonparametric statistical tests in this book are useful if a sample lacks normality.

2.5.1 Sample Kolmogorov–Smirnov One-Sample Test

A department store has decided to evaluate customer satisfaction. As part of a pilot study, the store provides customers with a survey to rate employee friendliness. The survey uses a scale of 1–10 and its developer indicates that the scores should conform to a normal distribution. Use the Kolmogorov–Smirnov one-sample test to decide if the sample of customers surveyed responded with scores approximately matching a normal distribution. The survey results are shown in Table 2.5.

TABLE 2.5

Survey results			
7	3	3	6
4	4	4	5
5	5	8	9
5	5	5	7
6	8	6	2

2.5.1.1 State the Null and Research Hypotheses

The null hypothesis states that the observed sample has an approximately normal distribution. The research hypothesis states that the observed sample does not approximately resemble a normal distribution.

The null hypothesis is

H_O: There is no difference between the observed distribution of survey scores and a normally distributed empirical sample.

The research hypothesis is

H_A: There is a difference between the observed distribution of survey scores and a normally distributed empirical sample.

2.5.1.2 Set the Level of Risk (or the Level of Significance) Associated with the Null Hypothesis

The level of risk, also called an alpha (α), is frequently set at 0.05. We will use an $\alpha = 0.05$ in our example. In other words, there is a 95% chance that any observed statistical difference will be real and not due to chance.

2.5.1.3 Choose the Appropriate Test Statistic

We are seeking to compare our observed sample against a normally distributed empirical sample. The Kolmogorov–Smirnov one-sample test will provide this comparison.

2.5.1.4 Compute the Test Statistic

First, determine the midpoint and standard deviation for the observed sample. Table 2.6 helps to manage the summations for this process.

TABLE 2.6

Survey score	Score frequency		
x_i	f_i	$f_i x_i$	$f_i x_i^2$
1	0	0	0
2	1	2	4
3	2	6	18
4	3	12	48
5	6	30	150
6	3	18	108
7	2	14	98
8	2	16	128
9	1	9	81
10	0	0	0
	$n = 20$	$\sum f_i x_i = 107$	$\sum f_i x_i^2 = 635$

Use Formula 2.11 to find the midpoint:

$$\begin{aligned} M &= (x_{\max} + x_{\min}) \div 2 \\ &= (9+2) \div 2 \\ M &= 5.5 \end{aligned}$$

Then, use Formula 2.12 to find the standard deviation:

$$\begin{aligned} s &= \sqrt{\frac{\sum (f_i x_i^2) - \frac{\left(\sum f_i x_i\right)^2}{n}}{n-1}} \\ &= \sqrt{\frac{635 - \frac{107^2}{20}}{20-1}} \\ s &= 1.81 \end{aligned}$$

Now, determine the z-scores, empirical relative frequencies, and observed relative frequencies for each score value (see Table 2.7).

TABLE 2.7

Survey score	Score frequency			Empirical frequency	Observed frequency
x_i	f_i	z-score	$\hat{p}_{x_i}$	$\hat{f}_r$	f_r
1	0	2.49	0.0064	0.006	0.000
2	1	1.93	0.0266	0.020	0.050
3	2	1.38	0.0838	0.064	0.100
4	**3**	**0.83**	**0.2033**	**0.140**	**0.150**
5	6	0.28	0.3897	0.250	0.300
6	3	0.28	0.3897	0.250	0.150
7	2	0.83	0.2033	0.140	0.100
8	2	1.38	0.0838	0.064	0.100
9	1	1.93	0.0266	0.020	0.050
10	0	2.49	0.0064	0.006	0.000

We will provide a sample calculation for survey score = 4 as seen in Table 2.7. Use Formula 2.13 to calculate the z-scores:

$$\begin{aligned} z &= \left| \frac{x_i - M}{s} \right| \\ &= \left| \frac{4 - 5.5}{1.81} \right| \\ z &= 0.83 \end{aligned}$$

Use each z-score and Table B.1 in Appendix B to determine the probability associated with the each value, $\hat{p}_{x_i}$:

$$\hat{p}_4 = 0.2033$$

To find the empirical frequency value $\hat{f}_r$ for each value, subtract its preceding value, $\hat{f}_{r-1}$, from the associated probability value $\hat{p}_{x_i}$. In other words,

$$\hat{f}_r = \hat{p}_{x_i} - \hat{f}_{r-1}$$

We establish our empirical frequency distribution beginning at the tail, $x_i = 1$, and work to the midpoint, $x_i = 5$:

$$\hat{f}_{r1} = \hat{p}_1 - \hat{f}_{r0} = 0.0064 - 0.000 = 0.006$$
$$\hat{f}_{r2} = \hat{p}_2 - \hat{f}_{r1} = 0.0266 - 0.006 = 0.020$$
$$\hat{f}_{r3} = \hat{p}_3 - \hat{f}_{r2} = 0.0838 - 0.020 = 0.064$$
$$\hat{f}_{r4} = \hat{p}_4 - \hat{f}_{r3} = 0.2033 - 0.064 = 0.140$$
$$\hat{f}_{r5} = \hat{p}_5 - \hat{f}_{r4} = 0.3897 - 0.140 = 0.250$$

Our empirical frequency distribution is based on a normal distribution, which is symmetrical. Therefore, we can complete our empirical frequency distribution by basing the remaining values on a symmetrical distribution. Those values are in Table 2.7.

Now, we find the values of the observed frequency distribution f_r with Formula 2.14. We provide a sample calculation with survey result = 4. That survey value occurs three times:

$$f_{r4} = \frac{f_{x_i=4}}{n} = \frac{3}{20}$$
$$f_r = 0.150$$

Next, we create cumulative frequency distributions using the empirical and observed frequency distributions. A cumulative frequency distribution is created by taking a frequency and adding all the preceding values. We demonstrate this in Table 2.8.

Now, we find the absolute value divergence $\tilde{D}$ and D between the cumulative frequency distributions. Use Formula 2.15 and Formula 2.16. See the sample calculation for survey score = 4 as seen in bold in Table 2.9.

$$\tilde{D}_4 = \left|\hat{F}_4 - S_4\right| = |0.230 - 0.300|$$
$$\tilde{D}_4 = 0.070$$

and

$$D_4 = \left|\hat{F}_4 - S_3\right| = |0.230 - 0.150|$$
$$D_4 = 0.080$$

TABLE 2.8

	Relative frequency		Cumulative frequency	
Survey score	Empirical	Observed	Empirical	Observed
x_i	$\hat{f}_r$	f_r	$\hat{F}_{x_i}$	S_{x_i}
1	0.006	0.000	0.006	0.000
2	0.020	0.050	$0.020 + 0.006 = 0.026$	$0.050 + 0.000 = 0.050$
3	0.064	0.100	$0.064 + 0.026 = 0.090$	$0.100 + 0.050 = 0.150$
4	0.140	0.150	$0.140 + 0.090 = 0.230$	$0.150 + 0.150 = 0.300$
5	0.250	0.300	$0.250 + 0.230 = 0.480$	$0.300 + 0.300 = 0.600$
6	0.250	0.150	$0.250 + 0.480 = 0.730$	$0.150 + 0.600 = 0.750$
7	0.140	0.100	$0.140 + 0.730 = 0.870$	$0.100 + 0.750 = 0.850$
8	0.064	0.100	$0.064 + 0.870 = 0.934$	$0.100 + 0.850 = 0.950$
9	0.020	0.050	$0.020 + 0.934 = 0.954$	$0.050 + 0.950 = 1.000$
10	0.006	0.000	$0.006 + 0.954 = 0.960$	$0.000 + 1.000 = 1.000$

TABLE 2.9

	Cumulative frequency		Cumulative frequency	
Survey score	Empirical	Observed	Divergence	
x_i	$\hat{F}_{x_i}$	S_{x_i}	$\tilde{D}$	D
1	0.006	0.000	0.006	
2	0.026	0.050	0.024	0.026
3	0.090	0.150	0.060	0.040
4	**0.230**	**0.300**	**0.070**	**0.080**
*5	0.480	0.600	0.120	*0.180
6	0.730	0.750	0.020	0.130
7	0.870	0.850	0.020	0.120
8	0.934	0.950	0.016	0.084
9	0.954	1.000	0.046	0.004
10	0.960	1.000	0.040	0.040

To find the test statistic Z, use the largest value from $\tilde{D}$ and D in Formula 2.17. Table 2.9 has an asterisk next to the largest divergence. That value is located at survey value = 5. It is $\max(|D|, |\tilde{D}|) = 0.180$:

$$\begin{aligned} Z &= \sqrt{n}\max(|D|, |\tilde{D}|) \\ &= \sqrt{20}(0.180) \\ Z &= 0.805 \end{aligned}$$

2.5.1.5 Determine the p-Value Associated with the Test Statistic The Kolmogorov–Smirnov test statistic Z and the Smirnov (1948) formula (see Formula 2.18, Formula 2.19, Formula 2.20, Formula 2.21, Formula 2.22, and Formula 2.23) are used to find the two-tailed probability estimate p. Since $0.27 \leq Z < 1$, we use Formula 2.19 and Formula 2.20:

$$\begin{aligned} Q &= e^{-1.233701Z^{-2}} \\ &= e^{-(1.233701)(0.805)^{-2}} \\ Q &= 0.149 \end{aligned}$$

and

$$\begin{aligned} p &= 1 - \frac{2.506628}{Z}(Q + Q^9 + Q^{25}) \\ &= 1 - \frac{2.506628}{0.805}(0.149 + 0.149^9 + 0.149^{25}) \\ p &= 0.536 \end{aligned}$$

2.5.1.6 Compare the p-Value with the Level of Risk (or the Level of Significance) Associated with the Null Hypothesis The critical value for rejecting the null hypothesis is $\alpha = 0.05$ and the obtained p-value is $p = 0.536$. If the critical value is greater than the obtained value, we must reject the null hypothesis. If the critical value is less than the obtained p-value, we must not reject the null hypothesis. Since the critical value is less than the obtained value ($0.05 < 0.536$), we do not reject the null hypothesis.

2.5.1.7 Interpret the Results We did not reject the null hypothesis, suggesting the customers' survey ratings of employee friendliness sufficiently resembled a normal distribution. This means that a parametric statistical procedure may be used with this sample.

2.5.1.8 Reporting the Results When reporting the results from the Kolmogorov–Smirnov one-sample test, we include the test statistic (D), the degrees of freedom (which equals the sample size), and the p-value in terms of the level of risk α. Based on our analysis, the sample of customers is approximately normal, where $D_{(20)} = 0.180$, $p > 0.05$.

2.5.2 Performing the Kolmogorov–Smirnov One-Sample Test Using SPSS

We will analyze the data from the example earlier using SPSS.

2.5.2.1 Define Your Variables First, click the "Variable View" tab at the bottom of your screen. Then, type the names of your variables in the "Name" column. As shown in Figure 2.13, the variable is called "Survey."

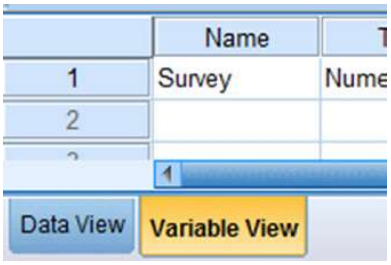

FIGURE 2.13

2.5.2.2 Type in Your Values Click the "Data View" tab at the bottom of your screen. Type your sample values in the "Survey" column as shown in Figure 2.14.

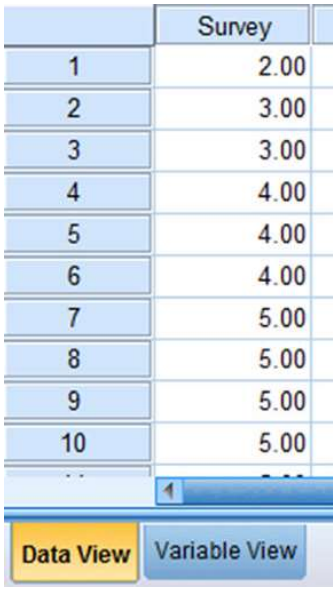

FIGURE 2.14

2.5.2.3 Analyze Your Data As shown in Figure 2.15, use the pull-down menus to choose "Analyze," "Nonparametric Tests," "Legacy Dialogs," and "1-Sample K-S . . ."

Use the arrow button to place your variable with your data values in the box labeled "Test Variable List:" as shown in Figure 2.16. Finally, click "OK" to perform the analysis.

2.5.2.4 Interpret the Results from the SPSS Output Window SPSS Output 2.2 provides the most extreme difference ($D = 0.176$), Kolmogorov–Smirnov Z-test statistic ($Z = 0.789$), and the significance ($p = 0.562$). Based on the results from SPSS, the p-value exceeds the level of risk associated with the null hypothesis ($\alpha = 0.05$). Therefore, we do not reject the null hypothesis. In other words, the sample distribution is sufficiently normal.

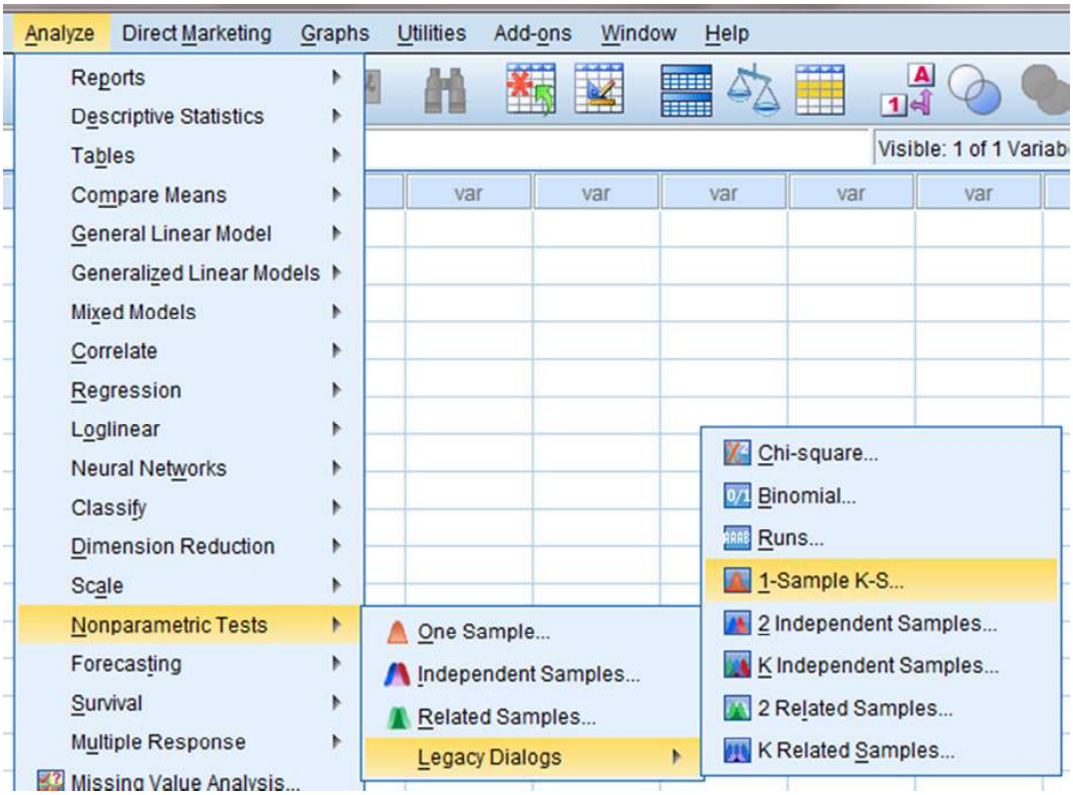

FIGURE 2.15

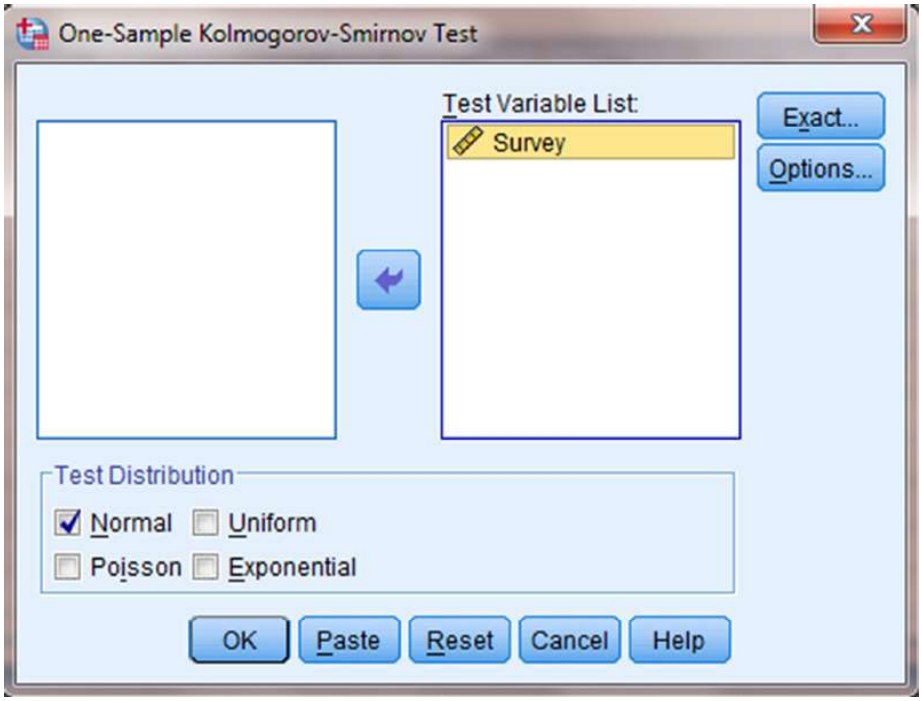

FIGURE 2.16

One-Sample Kolmogorov-Smirnov Test

		Survey
N		20
Normal Parameters[a,b]	Mean	5.3500
	Std. Deviation	1.81442
Most Extreme Differences	Absolute	.176
	Positive	.176
	Negative	-.124
Kolmogorov-Smirnov Z		.789
Asymp. Sig. (2-tailed)		.562

a. Test distribution is Normal.

b. Calculated from data.

SPSS OUTPUT 2.2

On an added note, differences between the values from the sample problem earlier and the SPSS output are likely due to value precision and computational round off errors.

2.6 SUMMARY

Parametric statistical tests, such as the *t*-test and one-way analysis of variance, are based on particular assumptions or parameters. Therefore, it is important that you examine collected data for its approximation to a normal distribution. Upon doing that, you can consider whether you will use a parametric or nonparametric test for analyzing your data.

In this chapter, we presented three quantitative measures of sample normality. First, we described how to examine a sample's kurtosis and skewness. Then, we described how to perform and interpret a Kolmogorov–Smirnov one-sample test. In the following chapters, we will describe several nonparametric procedures for analyzing data samples that do not meet the assumptions needed for parametric statistical tests. In the chapter that follows, we will begin by describing a test for comparing two unrelated samples.

2.7 PRACTICE QUESTIONS

1. The values in Table 2.10 are a sample of reading-level score for a 9th-grade class. They are measured on a ratio scale. Examine the sample's skewness and kurtosis for normality for $\alpha = 0.05$. Report your findings.
2. Using a Kolmogorov–Smirnov one-sample test, examine the sample of values from Table 2.10. Report your findings.

TABLE 2.10

Ninth-grade reading-level score									
8.10	8.20	8.20	8.70	8.70	8.80	8.80	8.90	8.90	8.90
9.20	9.20	9.20	9.30	9.30	9.30	9.40	9.40	9.40	9.40
9.50	9.50	9.50	9.50	9.60	9.60	9.60	9.70	9.70	9.90

2.8 SOLUTIONS TO PRACTICE QUESTIONS

1. SPSS returned the following values:

skewness $= -0.904$

standard error of the skewness $= 0.427$

kurtosis $= 0.188$

standard error of the kurtosis $= 0.833$

The computed z-scores are as follows:

$$z_{S_k} = -2.117$$

and

$$z_K = 0.226$$

At $\alpha = 0.05$, the sample's skewness fails the normality test, while the kurtosis passes the normality test. Based on our standard of $\alpha = 0.05$, this sample of reading levels for 9th-grade students is not sufficiently normal.

2. SPSS Output 2.3 shows the results from the Kolmogorov–Smirnov one-sample test.
Kolmogorov–Smirnov obtained value $= 1.007$
Two-Tailed significance $= 0.263$

One-Sample Kolmogorov-Smirnov Test

		Scores
N		30
Normal Parameters[a,b]	Mean	9.1800
	Std. Deviation	.46639
Most Extreme Differences	Absolute	.184
	Positive	.099
	Negative	-.184
Kolmogorov-Smirnov Z		1.007
Asymp. Sig. (2-tailed)		.263

a. Test distribution is Normal.

b. Calculated from data.

SPSS OUTPUT 2.3

According to the Kolmogorov–Smirnov one-sample test with $\alpha = 0.05$, this sample of reading levels for 9th-grade students is sufficiently normal.