# C H A P T E R   3

# Evaluating the Characteristics of Data

**C**hapter 2 focused on the process of statistical hypothesis testing. Part of this process (Step 6) involves evaluating the extent to which the data being analyzed meet the assumptions of the tests being considered. Chapter 3 will outline available methods for evaluating the characteristics of data. First, the level of measurement of a variable needs to be identified to determine the most appropriate parametric or nonparametric statistical test. Next, it is important to evaluate the normality of the variable's distribution, the impact of outliers, the homogeneity of variance, and sample size adequacy.

## CHARACTERISTICS OF LEVELS OF MEASUREMENT

Measurement is the process of assigning numbers or codes to observations according to certain prescribed rules. The way in which these values are assigned to the observations determines a variable's *level of measurement*. The most widely accepted set of rules for determining a variable's level of measurement is that developed by S. Stevens (1946). This typology consists of four levels of measurement whose order is based on how much information they carry. These levels are nominal, ordinal, interval, and ratio. Table 3.1 summarizes the characteristics of these four levels of measurement.

### Nominal

The first level of measurement is nominal. A variable that is measured on a nominal scale is one that has distinct nonoverlapping categories. The numbers that are assigned to these categories have no intrinsic meaning, but all persons who share the same category are assigned a similar value.

There are three basic requirements for a "good" nominal-level variable: (1) all members of one level of the variable must be assigned the same number, (2) no two levels are assigned the same number, and (3) each observation can be assigned to one and only one of the available levels. Given that these three conditions have been fulfilled, the levels of the nominal-level variable are mutually exclusive and exhaustive.

**17**

| Table 3.1 | Overview of the Characteristics of the Levels of Measurement |
|---|---|

| Level of Measurement | Mutually Exclusive Groups | Rank Ordering | Equidistant Values | Meaningful Zero Point | Example |
|---|---|---|---|---|---|
| Nominal | • | | | | Marital status |
| Ordinal | • | • | | | Stress level (1–7) |
| Interval | • | • | • | | Depression scale (1–100) |
| Ratio | • | • | • | • | Weight (pounds) |

The variable *gender* is a nominal-level measurement because it is composed of two independent, mutually exclusive (nonoverlapping), and exhaustive levels: male and female. In our hypothetical intervention study, each of the 20 participating children could be assigned a "0" or a "1" depending on whether the child is a male (0) or a female (1). The numbers 0 and 1 that have been assigned to these levels have no inherent order to them; these numbers could have been reversed. They merely indicate the gender group to which the child belongs. Additional variables in our hypothetical study that have a nominal level of measurement are the group to which the child was assigned (intervention = 0, usual care = 1), diagnosis (1 = solid tumor, 2 = acute myeloid leukemia, 3 = lymphoma, 4 = sarcoma), and race/ethnicity (1 = Caucasian, 2 = African American, 3 = Hispanic or Latino, and 4 = other).

Parametric statistics assume ordering and meaningful numerical distances between values; therefore, these statistics do not provide very useful information if the dependent or outcome variable has a nominal level of measurement. It does not make sense, for example, to report an average marital status. For nominal data, researchers rely instead on frequencies, percentages, and modes to describe their results. Nonparametric inferential statistics (e.g., the chi-square goodness-of-fit test or Fisher's exact test) may also be applied to these data.

## Ordinal

The next level of measurement is ordinal. A variable that has an ordinal level of measurement is characterized by having mutually exclusive categories that are sorted and rank ordered on the basis of their standing relative to one another on a specific attribute according to some preset criteria. Although it may be possible to ascertain that one person has a higher rank relative to another person, it is not possible to determine exactly how much higher that person is than another.

Suppose the nurses in our hypothetical intervention study were asked to assess on a 7-point scale (1 = *not at all distressed* to 7 = *very distressed*) the extent to which a particular child appears to be distressed prior to our planned intervention. This

variable, preintervention distress, is an ordinal-level variable. We know, for example, that Child A, who received a "6" on preintervention distress, was more distressed prior to the intervention than Child B, who received a "3" on this scale. Because there are not equidistant intervals on this 7-point scale, however, it is not possible to conclude that Child A is twice as distressed as Child B or that the difference between a "6" and a "7" is the same as the difference between a "3" and a "4." Moreover, not all values necessarily share the same intensity. For example, Nurse C's assignment of a "7" to a child may not have the same intensity level as Nurse D's "7." We only know that, for both nurses, a particular child was "very distressed" according to their criteria.

Because there is order to the values of an ordinal scale, descriptive statistics that rely on rank ordering (e.g., the median) can be used in addition to percentages, frequencies, and modes. Numerous nonparametric inferential statistics are available to test hypotheses about similarities of medians between groups and relationships among variables.

There has been much heated discussion in the research literature about the appropriateness of using parametric tests with ordinal-level data (Armstrong, 1981; Carifio & Perla, 2008; Jamieson, 2004; Knapp, 1990; Norman, 2010; Pell, 2005). Pedhazur and Schmelkin (1991) suggest that this controversy was sparked by early writings of S. Stevens (1951), who argued that means and standard deviations, the backbones of parametric statistics, were not appropriate measures of central tendency for ordinal data. Others have effectively argued (Knapp, 1990) that the critical issue is not so much that the data are ordinal but rather that the data have a sufficient sample size (e.g., $N > 30$) and a relatively normal distribution of the dependent variable to merit the use of parametric statistics. Norman (2010) presents a convincing argument that parametric statistics can be used with Likert data even with small sample sizes, unequal variances, and nonnormal distributions.

## Interval

Interval-level scales are more refined than either nominal or ordinal scales. Like the ordinal scale, the interval-level scale has mutually exclusive groups and rank ordering. Unlike the ordinal scale, the interval-level scale has equidistant intervals. This means that we obtain information not only about the rank order of a particular score but also about how much greater or less a particular score is than another. That is, on an interval scale whose range is 1 to 100, the difference between 100 and 75 is, in some sense, the same as the difference between 75 and 50.

A classic example of an interval-level scale is temperature measured in degrees Fahrenheit. We know, for example, that a child whose body temperature is 102° has a temperature that is 2° higher than a child whose body temperature is 100°. Because an interval-level scale does not have an absolute zero point, however, the distances between values, although theoretically equidistant, do not carry exactly the same meaning. That is, the change in body temperature from 98° to 101° is not meaningfully the same as a change in body temperature from 102° to 105°. However, 100° is not twice as hot as 50° because 0° Fahrenheit is a numerical convenience, not an absolute.

A common practice among researchers is to use a multi-item scale to measure single or multiple constructs. The individual items tend to be either nominal (e.g., 0 = *agree* vs. 1 = *disagree*) or ordinal (e.g., 1 = *strongly agree* to 5 = *strongly disagree*) in nature, and the item responses are summed to produce a scale with interval-level properties and with a larger range of possible scores (e.g., 0–100). From these data, we can use all the measures of central tendency and variance. Parametric statistics such as the *t* test, analysis of variance (ANOVA), and Pearson product-moment correlation coefficient are all possible considerations.

In our intervention example, we might decide to use a 14-item self-reported fatigue assessment scale for children ages 7 to 12 years (Hinds et al., 2007; Hockenberry et al., 2003). This Childhood Fatigue Scale (CFS) is a 14-item instrument that first asks the child for a "yes" or "no" response regarding their experiences of 14 fatigue-related symptoms (e.g., *I have been tired*). If the child answers yes to the symptom, he or she is then asked to describe the intensity of the fatigue symptom on a scale of 1 (*not at all*) to 5 (*a lot*). From these 14 items, a total fatigue score can be generated with a range of scores from 0 (no fatigue) to 70 (high fatigue) along with three subscales: lack of energy, inability to function, and altered mood. (Hinds & Hockenberry-Eaton, 2001; Hockenberry et al., 2003).

Again, controversy exists as to the true nature of the level of measurement of such a multi-item scale (Knapp, 1990; Nunnally & Bernstein, 1994; Pedhazur & Schmelkin, 1991). That is, is an "interval" scale that has been generated from ordinal data truly interval? Should we even care? For statistical analysis, the concern is not so much the variable's "true" level of measurement as much as whether the information generated from the use of a particular statistic best represents the data. This conclusion can be reached only by examining the data thoroughly to determine the extent to which a particular test's assumptions have been violated. Pedhazur and Schmelkin (1991) indicate that, even in his later writings, S. Stevens (1968) argued, "The question is thereby made to turn, not on whether the measurement scale determines the choice of a statistical procedure, but on how and to what degree an inappropriate statistic may lead to a deviant conclusion" (p. 852).

## Ratio

The highest level of measurement is ratio. In addition to maintaining the character-istics of the previous three levels of measurement (mutually exclusive and exhaustive categories, rank ordering, and equidistant intervals), a ratio-level variable also has a meaningful and absolute zero point that represents the complete absence of a given attribute. Because of its invariant zero point, the ratio of any two scores from a ratio scale is unchanged by transformations through multiplication and division.

Examples of ratio-level variables include weight, blood pressure, and temperature Kelvin. In our hypothetical study, a child's body weight and time to first voiding could be considered ratio-level variables. The age of the child might be more controversial. Our society has yet to agree on when an individual becomes a human being. At conception? At birth? Or at some other place along the way?

It does not matter much in statistics whether a variable is at the interval or ratio level of measurement. Both of these levels of measurement are appropriate for use with parametric statistics. To reiterate, equally important determinations regarding the use of parametric statistics are sample size and the shape of the distribution of the dependent variable.

## Which Level of Measurement Is "Best"?

There is no clear answer as to which level of measurement is best for a particular research question. Clearly, the researcher wants to attain the very highest level of measurement possible given the time, financial, and design constraints of the research. The higher levels of measurement, interval and ratio, provide the researcher with the opportunity to use potentially more powerful statistical tests. Moreover, it is always possible to collapse data into lower levels of measurement. It is not possible, however, to resurrect interval-level data from precollapsed nominal data. The best approach is *not* to collapse data while entering them into the computer. Data can be collapsed, if necessary, later on during the statistical analyses.

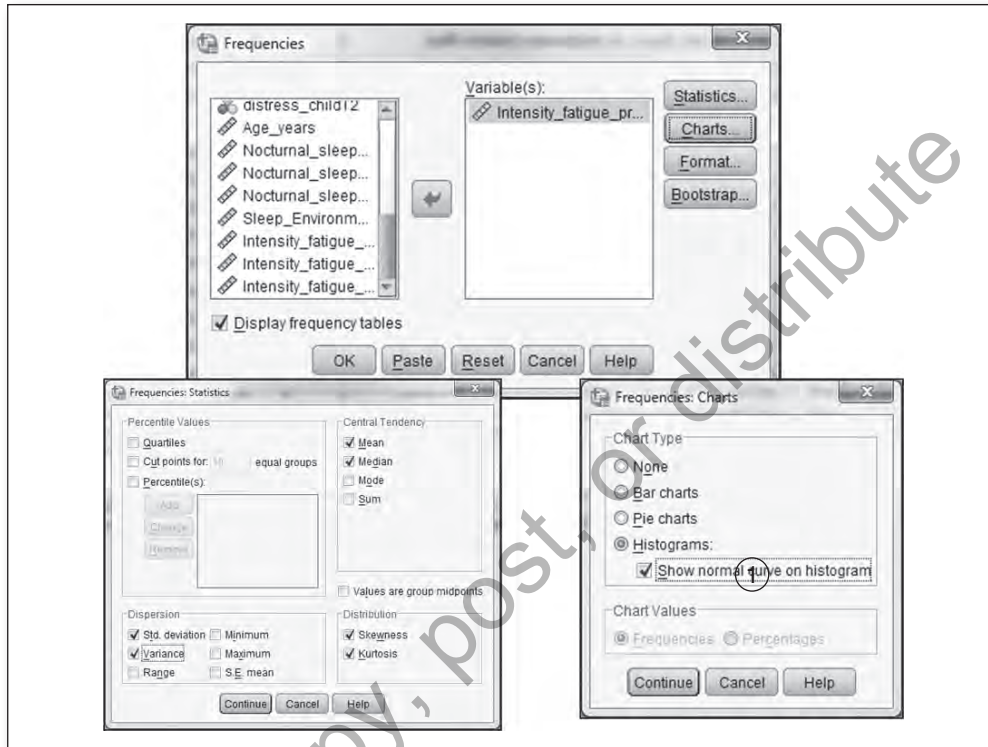## ASSESSING THE NORMALITY OF A DISTRIBUTION

Returning to our hypothetical intervention study, suppose that we were interested in assessing the normality of the distribution of scores for children's self-reported fatigue during the 24 hours prior to the implementation of our intervention. As indicated above, this is a variable whose scores can range from 0 to 70, with higher scores suggesting greater intensity of fatigue. There are several ways that we could assess the normality of this variable. First, we could examine the distribution's skewness and kurtosis. Next, we could visually examine the distribution of the data to obtain a sense of its shape. Finally, we could statistically test the extent to which the data fit a theoretically normal distribution.

All three of these approaches are available in SPSS for Windows by choosing the following commands from the dropdown menu: (a) *Analyze . . . Descriptive Statistics . . . Frequencies . . .* (Figure 3.1) and (b) *Analyze . . . Descriptive Statistics . . . Explore . . .* (Figure 3.2). The *Frequencies* and *Explore* dialog boxes allow the researcher a number of options for evaluating data.

As indicated in Figure 3.1, by opening the *Frequencies . . . Charts* dialogue box and selecting *Histograms . . . with normal curve,* a normal distribution can be superimposed over the histogram of the variable of interest ①. This allows the researcher to visually inspect the data for violations of normality. The *Analyze . . . Descriptive Statistics . . . Explore* command may also be used to statistically test for normality (Figure 3.2, ①). This procedure also produces information regarding descriptive statistics, stem-and-leaf plots, boxplots, outliers, normal probability plots, and statistical tests of normality. Separate analyses can be obtained for subgroups of data as well.

**Figure 3.1** SPSS for Windows *Analyze . . . Descriptive Statistics . . . Frequencies . . .* commands for assessing normality of a distribution.
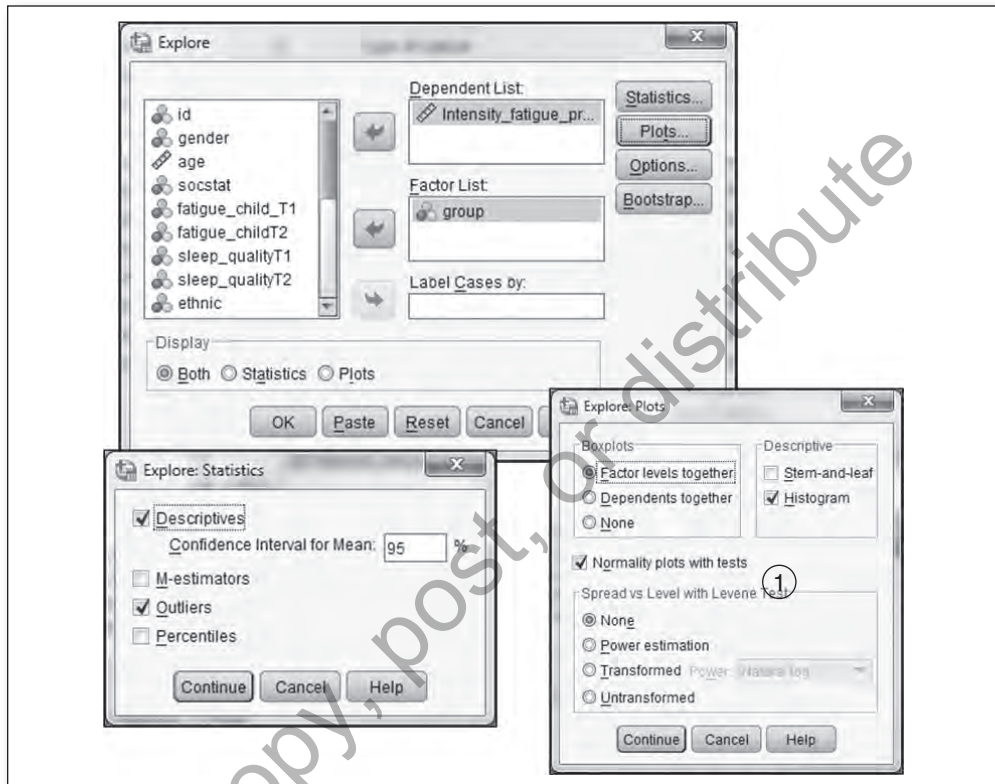
## Skewness

Before we interpret the results of our SPSS output, let's review the meaning of skewness and kurtosis. You will recall that a normal distribution takes the form of a bell-shaped curve that is centered on the mean (Figure 3.3A). The normal distribution is symmetric, and all three measures of central tendency—the mean, median, and mode—share the same value. One simple way of assessing normality of a distribution, therefore, is to examine the measures of central tendency. If the mean, median, and mode are nearly equal in value, then there is evidence to suggest that the distribution is symmetric. If these three values are not at all similar, then the distribution is characterized as being asymmetric or *skewed;* that is, the distribution has one tail that is longer than the other.

There are two types of skewness: positive and negative skew. A distribution is *positively skewed* if the distribution's longer tail extends toward the right or toward the higher set
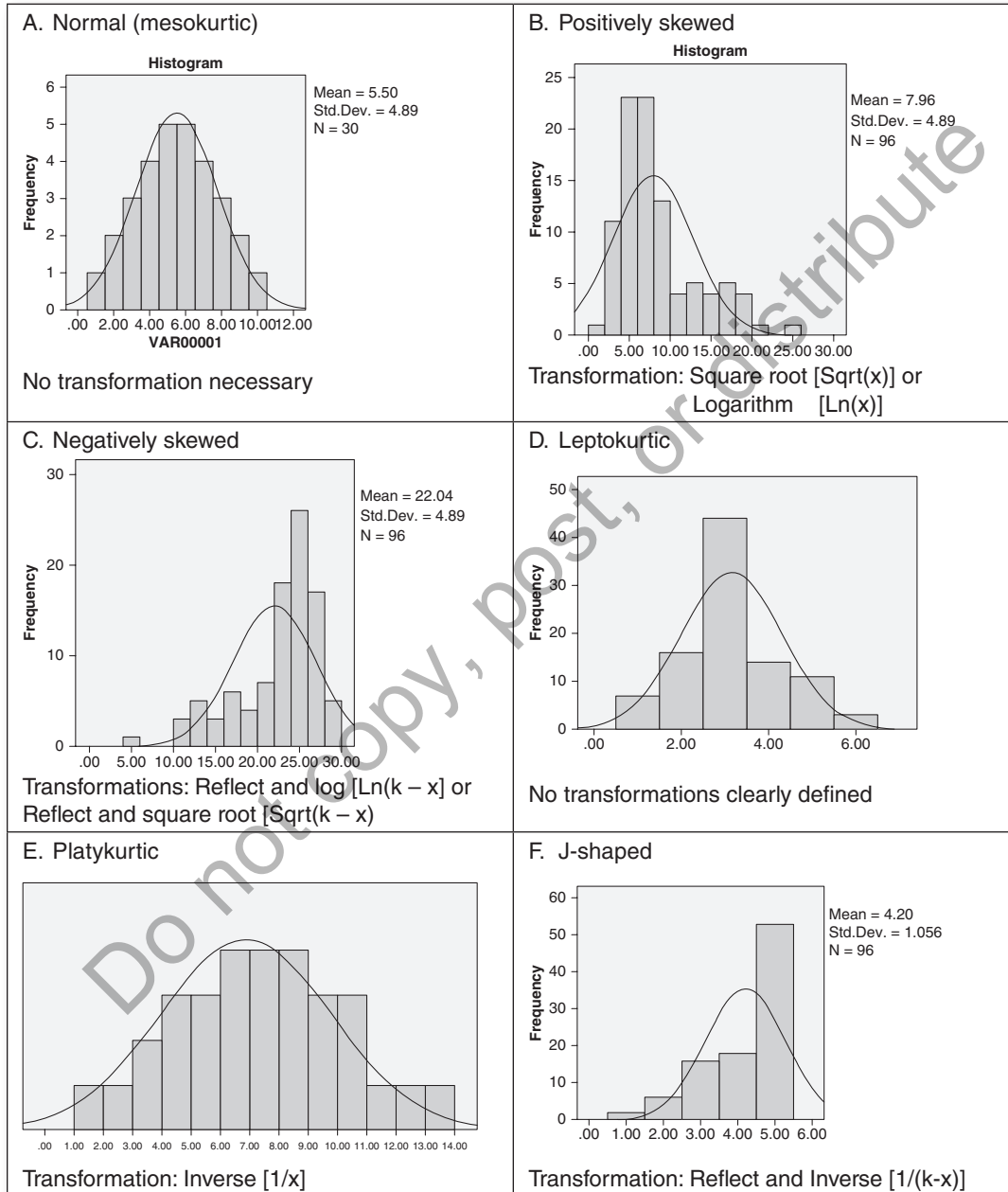
of values (Figure 3.3B). This results in the value of the mean being pulled to the right and being larger in value than the median or mode. A distribution that is *negatively skewed* has a longer tail that extends toward the left or toward the lower set of values (Figure 3.3C). This results in the mean being smaller in value than the median or mode.

The measure of skewness is referred to as the third moment about the mean, $\Sigma(X - \overline{X})^3/(n - 1)$ (Neter, Wasserman, & Whitmore, 1993; Park, 2008). Because this third moment is measured in cubed units (e.g., weight cubed), a standardized measure of skewness is considered more useful because its size does not depend on the units of measurement. This standardized measure is obtained by dividing the third moment by the cube of the standard deviation ($s^3$) of the variable being examined (Neter et al., 1993). This is the skewness value that is presented in the computer printout.

When a distribution is a symmetric bell-shaped curve, the value of this measure of skewness is 0. The measure has a negative value when the distribution is negatively

| Figure 3.3 | Comparison of the most common forms of distributions and suggested transformations. |
|---|---|

**A. Normal (mesokurtic)**

**Histogram**

Mean = 5.50
Std.Dev. = 4.89
N = 30

No transformation necessary

**B. Positively skewed**

**Histogram**

Mean = 7.96
Std.Dev. = 4.89
N = 96

Transformation: Square root [Sqrt(x)] or
Logarithm    [Ln(x)]

**C. Negatively skewed**

Mean = 22.04
Std.Dev. = 4.89
N = 96

Transformations: Reflect and log [Ln(k − x] or
Reflect and square root [Sqrt(k − x)

**D. Leptokurtic**

No transformations clearly defined

**E. Platykurtic**

Transformation: Inverse [1/x]

**F. J-shaped**

Mean = 4.20
Std.Dev. = 1.056
N = 96

Transformation: Reflect and Inverse [1/(k-x)]

SOURCE: Transformation suggestions come from Hair, Anderson, Tatham, and Black (1995) and Tabachnick and Fidell (2013).
In Panels C and F, *k* represents a constant, usually the largest score +1.

skewed and a positive value when the distribution is positively skewed. To determine the seriousness of the skewness of a distribution, one of two measures of skewness, Fisher's or Pearson's, can be used (Kellar & Kelvin, 2012; Lehman, 1991; Salkind, 2010). The Fisher's coefficient is as follows:

$$\text{Fisher skewness coefficient} = \text{skewness} / \text{standard error skewness}.$$

To calculate the Fisher skewness coefficient, the SPSS computer-generated value for skewness (Skewness) is divided by the standard error for skewness (SE Skew). If the resulting $z$ statistic lies beyond the range of $\pm 1.96$ (the critical value for a two-tailed $z$ statistic at $\alpha = .05$), the distribution is asymmetric and significantly skewed. Calculated values of this coefficient that fall between $-1.96$ and $+1.96$ suggest that the distribution is not significantly different from a normal distribution.

A second commonly used index for skewness is the Pearson skewness coefficient ($Sk_p$):

$$\text{Pearson skewness coefficient} = Sk_p = 3[(\overline{X} - Md) / s].$$

This statistic uses the difference between the mean ($\overline{X}$) and median (Md) of a distribution divided by the variable's standard deviation ($s$) to determine the level of skewness. If $Sk_p = 0$, the mean and median are equal and therefore the distribution is symmetric. A negative coefficient indicates a negative skew (i.e., the mean is smaller than the median), and a positive value represents positive skewness (i.e., the mean is larger than the median). Lehman (1991) suggests that values of $Sk_p$ between $-0.5$ and $+0.5$ indicate generally acceptable levels of skewness.

## Kurtosis

A second characteristic of a distribution is its *kurtosis,* or the fourth moment about the mean (Balanda & MacGillivray, 1988; DeCarlo, 1997; Neter et al., 1993; Park, 2008), calculated as $\Sigma(X - \overline{X})^4 / (n - 1)$. Because this fourth moment is measured in (units)$^4$, a standardized measurement of kurtosis is available in statistical packages that divides the fourth movement by $s^4$. Evaluation of a distribution's kurtosis is especially useful after it has been determined that the distribution is not unduly skewed. It is not very useful for asymmetric or skewed distributions.

Three terms are used to denote different levels of kurtosis: *mesokurtic, leptokurtic,* and *platykurtic*. A normal distribution has a standardized kurtosis value that is equal to zero and is referred to as being *mesokurtic* (Figure 3.3A). A positive value for the standardized kurtosis coefficient implies that the distribution is *leptokurtic,* or more peaked than a normal distribution (Figure 3.3D). (To remember what *leptokurtic* means, it might be helpful to recall Superman, *leaping* tall buildings in a single bound.) A negative value for the standardized kurtosis coefficient implies that the distribution is *platykurtic,* or flatter than a normal distribution (Figure 3.3E). (Remember that, like the platykurtic distribution, a platypus is an animal that stands low and close to the ground.) Kellar and Kelvin (2012) suggest using a Fisher coefficient to evaluate kurtosis:

$$\text{Fisher coefficient of kurtosis} = \text{kurtosis} / \text{standard error of kurtosis}.$$

That is, the standardized kurtosis value is divided by its standard error to determine the extent to which a bell-shaped symmetric curve deviates from a normal distribution. If this $z$ statistic falls outside the range of ±1.96, then the bell-shaped distribution is significantly different from a standard normal distribution.

## Computer Analysis of Skewness and Kurtosis

Assuming that the data for the 20 subjects in our hypothetical study have been entered into the computer data file (we are using "hospitalized children with cancer-20 cases.sav" located on the SAGE website, study.sagepub.com/pett2e), the syntax commands and computer-generated frequency output for the child's self-reported fatigue preintervention are presented in Figure 3.4. This output was obtained by running the commands presented in Figure 3.1 and selecting the preintervention fatigue variable for analysis.

We are first presented with the frequency distribution of the child's preintervention fatigue variable (Figure 3.4, (1)). Recall that the scores could range from 0 to 70, with higher scores indicating greater fatigue. Given that no child indicated that he or she was "very fatigued" during the preintervention phase, there are no values higher than 35. Notice that the mean (29.25), median (30.0), and mode (35.0) (2), although close, are not equal to one another, suggesting that the data may be skewed. Because the mean is smaller than either the median or the mode, the data are negatively skewed, with the longer tail in the direction of the smaller values, a condition that is verified by the skewness value of −1.103 (3). Dividing the measure of skewness by the standard error for skewness (−1.103/.512) results in a Fisher skewness coefficient of −2.15, which falls outside the acceptable limits of ±1.96, suggesting that the data may be seriously skewed. It is interesting that a different result is obtained for the Pearson $Sk_p$:

$$Sk_p = 3(\overline{X} - Md)\,/\,s = 3(29.25 - 30.0)\,/\,(6.54) = -.34.$$

The resulting value of −.34 is within the acceptable range of this coefficient (−.5 to + .5). This discrepancy may be explained by the extreme sensitivity of the Fisher measure of skewness to outliers (Kellar & Kelvin, 2012). Because the statistic is based on deviations from the mean raised to the third power, outliers have a very strong effect on this measure.
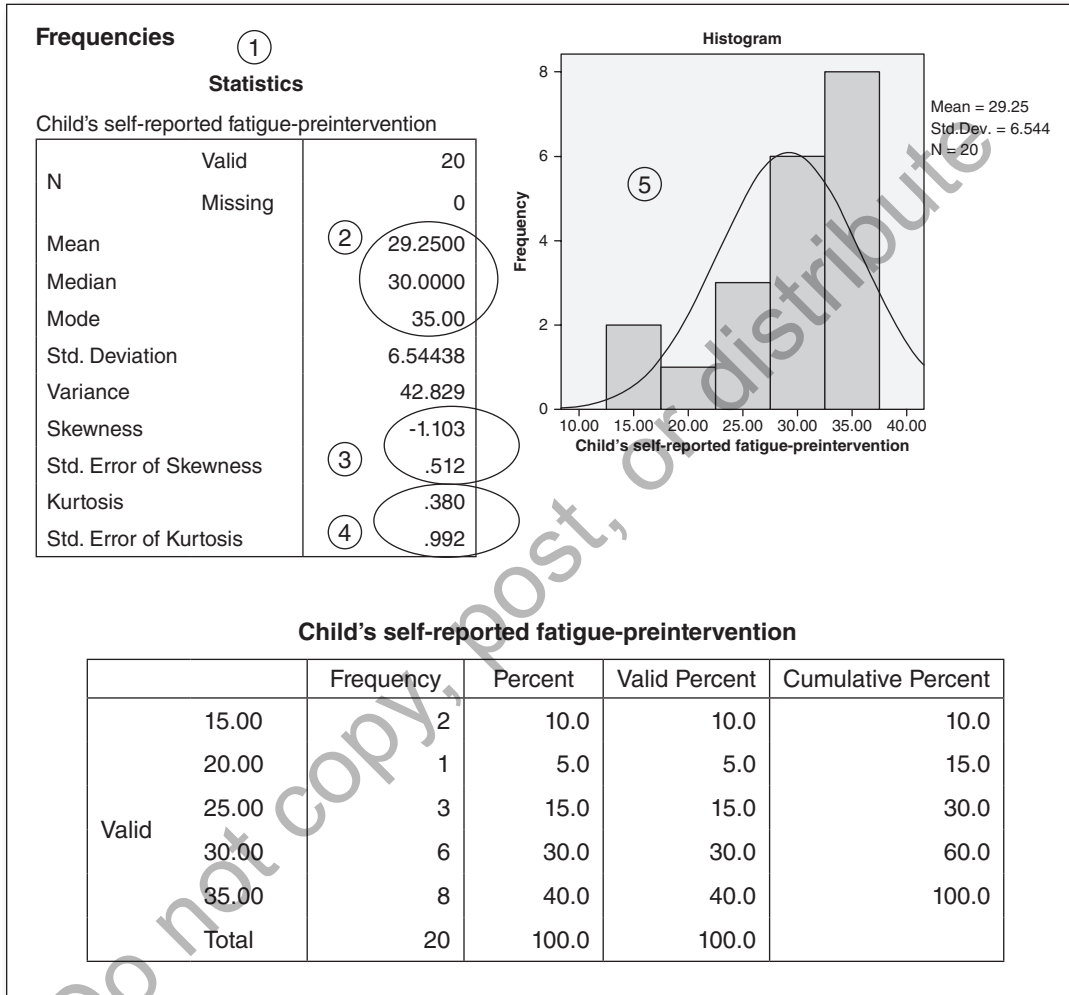
Ordinarily, when a distribution has serious skewness problems, indicating that it is not bell-shaped, it would not be necessary to examine its kurtosis. Given that we have conflicting information regarding this distribution's skewness, however, it would be useful to examine the distribution's kurtosis as well. The positive value (.380) for kurtosis (4) indicates that the distribution is leptokurtic, or more peaked than a normal distribution. Dividing the value for kurtosis by its standard error (.992), however, yields a Fisher coefficient of kurtosis of .383, which is well within the ±1.96 range for a normal distribution.

## Visually Examining the Shape of the Distribution

Given these somewhat conflicting results, it is important that we examine the data visually to determine for ourselves the seriousness of the skewness. In fact, the

## Figure 3.4

Computer-generated output obtained in SPSS for Windows (v. 22–23) for the frequencies and histogram of preintervention fatigue.

**Frequencies** ①

**Statistics**

Child's self-reported fatigue-preintervention

| | | |
|---|---|---|
| N | Valid | 20 |
| | Missing | 0 |
| Mean | ② | 29.2500 |
| Median | | 30.0000 |
| Mode | | 35.00 |
| Std. Deviation | | 6.54438 |
| Variance | | 42.829 |
| Skewness | | -1.103 |
| Std. Error of Skewness | ③ | .512 |
| Kurtosis | | .380 |
| Std. Error of Kurtosis | ④ | .992 |

**Histogram**

Mean = 29.25
Std.Dev. = 6.544
N = 20

⑤

Child's self-reported fatigue-preintervention

**Child's self-reported fatigue-preintervention**

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | 15.00 | 2 | 10.0 | 10.0 | 10.0 |
| | 20.00 | 1 | 5.0 | 5.0 | 15.0 |
| | 25.00 | 3 | 15.0 | 15.0 | 30.0 |
| | 30.00 | 6 | 30.0 | 30.0 | 60.0 |
| | 35.00 | 8 | 40.0 | 40.0 | 100.0 |
| | Total | 20 | 100.0 | 100.0 | |

necessity of visually examining data for departures from normality cannot be over-stressed. No statistical test of normality is superior to what my biostatistician friend, Dr. James Reading, refers to as the *eyeball test*.

The eyeball test consists of visually examining the data's distribution to determine if the distribution looks sufficiently comparable to a normal distribution for the researcher to feel comfortable using parametric tests. Is the mean an adequate

representation of these data? Are there unusual "kinks" in the distribution? Is the distribution unimodal, or are there multiple modes? Are there outliers about which to be concerned? What effect does the sample size have on the potential shape of the distribution? Although the mean, median, and mode may be similar, a limited sample size may restrict one's ability to adequately distinguish the shape of the distribution. If the data do not have a normal distribution, is there a possible transformation that could be performed (e.g., log or square root) that would make sense logically and that would transform the nonnormal distribution into a more nearly normal distribution?

Figure 3.4 also presents a graph of the normal curve superimposed on the distribution for the preintervention fatigue variable for the 20 subjects in our hypothetical study ⑤. This figure indicates that the data are negatively skewed and somewhat leptokurtic in shape. The distribution also appears to have serious deviations from normality. With so few data points ($N = 20$), the shape of the distribution may also not be definitively determined. This information suggests that the use of nonparametric tests with these data may be in order. A second alternative would be to consider the possibility of transforming the preintervention fatigue variable to obtain a more nearly normal distribution.

Additional plots of normality may be generated in SPSS for Windows (v. 22–23) through the *Analyze . . . Descriptive Statistics . . . Explore . . . Plots . . . Normality plots with tests . . .* commands. These plots can be of help in visually examining data. They include normal probability plots and detrended normal plots. Figure 3.5 presents examples of normal and detrended normal probability plots for selected distributions.
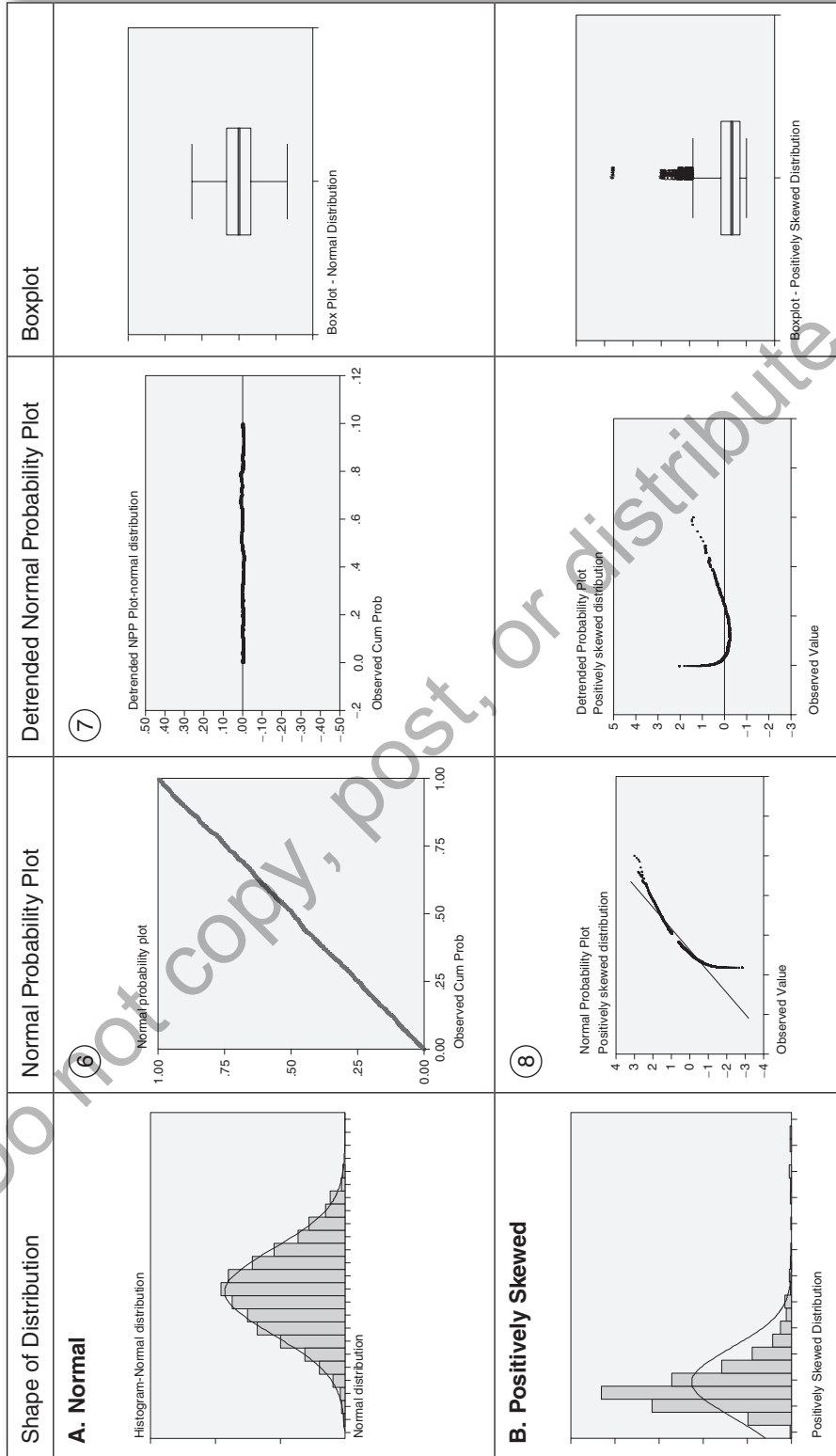
### Normal and Detrended Normal Probability Plots.

In the *normal probability plot*, each data point is paired with its expected value given a nearly normal distribution of similar range and sample size. If the sample is from a nearly normal distribution (Figure 3.5A), a normal probability plot of the observed and expected values would indicate that nearly all values lie along a 45° straight line running from the lower left corner to the upper right corner of the plot ⑥. Note that, except for a few minor deviations, the values fall along the 45° line.

A *detrended normal probability plot* is one in which the deviations from normal for each value in the sample are plotted against the observed values. If the sample is from a nearly normal distribution, these deviations will cluster evenly around zero along a horizontal band. This indicates that there is little difference between the observed values and expected values. The detrended normal probability plot for the nearly normal distribution in Figure 3.5A, third panel ⑦, illustrates this pattern. Note that the data do *not* need to fall *exactly* along a straight line but rather that the band of values is similar in width across all values of the data.
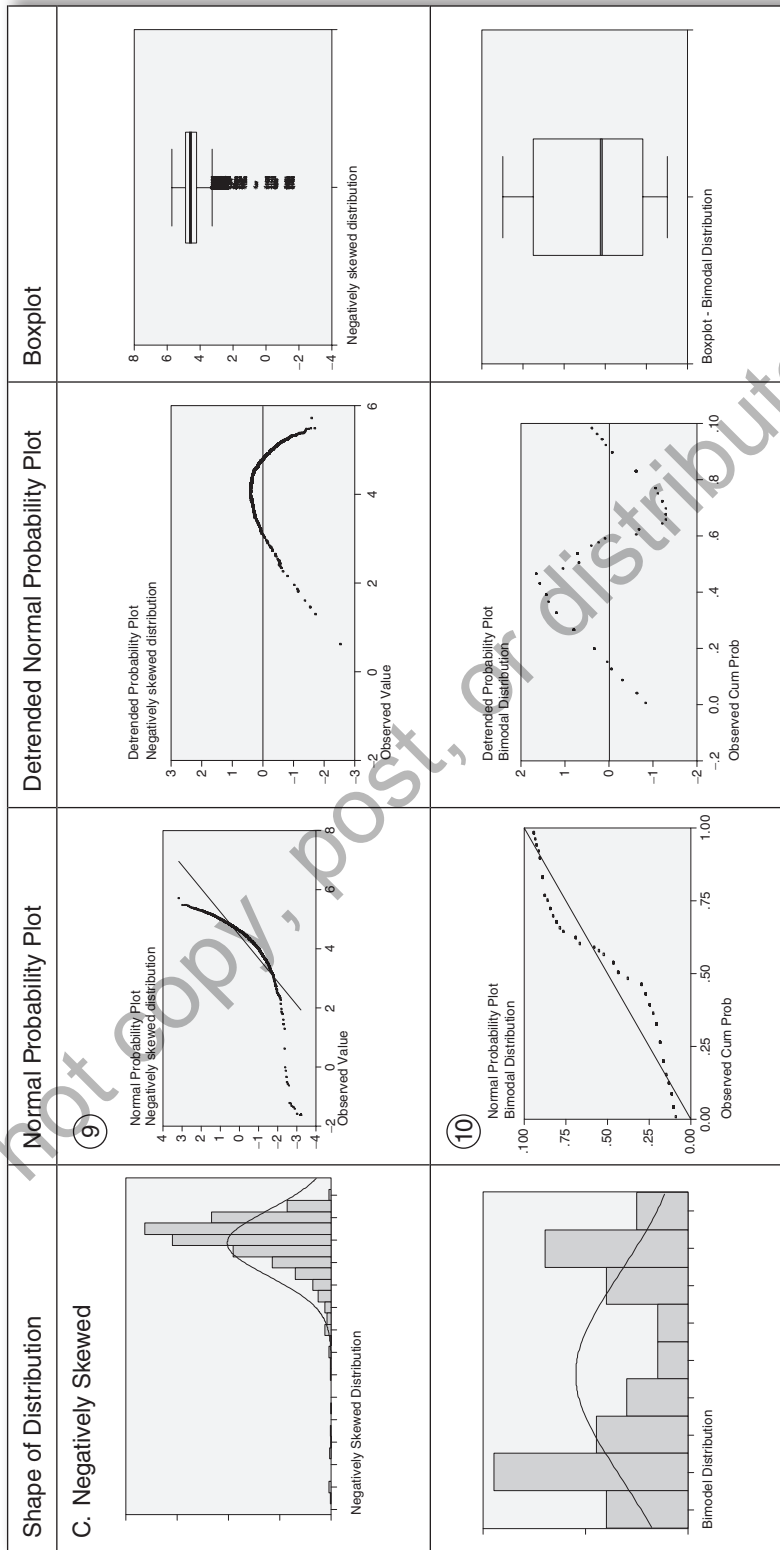
Distributions that are skewed or bimodal (e.g., Figures 3.5B–D) show markedly different patterns of deviations from normality. Curvilinear patterns often emerge, suggesting that the data are badly skewed (Figures 3.5B–C) ⑧ ⑨ or bimodal (Figure 3.5D) ⑩. Outliers can be identified on these plots because they occupy

**Figure 3.5** Examples of normal probability, detrended normal probability, and boxplots for the normal and other selected distributions.

| Shape of Distribution | Normal Probability Plot | Detrended Normal Probability Plot | Boxplot |
|---|---|---|---|

**A. Normal**



*(Continued)*

**29**

**Figure 3.5** (Continued)

| Shape of Distribution | Normal Probability Plot | Detrended Normal Probability Plot | Boxplot |
|---|---|---|---|
| C. Negatively Skewed | | | |



Negatively Skewed Distribution

⑨ Normal Probability Plot
Negatively skewed distribution

Observed Value

Detrended Probability Plot
Negatively skewed distribution

Observed Value

Negatively skewed distribution



Bimodal Distribution

⑩ Normal Probability Plot
Bimodal Distribution

Observed Cum Prob

Detrended Probability Plot
Bimodal Distribution

Observed Cum Prob

Boxplot - Bimodal Distribution

positions away from the other values and do not appear to be connected to them (Tabachnick & Fidell, 2013). For example, in Figure 3.5C ⑨, the values in the lower left-hand corner of the normal and detrended normal probability plots represent the outliers for this negatively skewed distribution.

### Computer Examples of the Plots.

Figure 3.6 presents normal probability and detrended normal probability plots for the self-reported preintervention fatigue variable. The plots were generated from the SPSS for Windows (v. 22–23) commands *Analyze . . . Descriptive Statistics . . . Explore . . . Normality plots with tests . . .* illustrated in Figure 3.2. The data file that we are using is hospitalized children with cancer-20 cases.sav found on the SAGE website, study.sage pub.com/pett2e.

The two plots in Figure 3.6 confirm what we saw in Figure 3.4, that the preintervention fatigue data are not normally distributed. The values for this fatigue variable are not similar to the expected values and, therefore, are not situated on the 45° straight line of the normal probability plot (Figure 3.6A) ⑪. The detrended plot (Figure 3.6B) ⑫ indicates that the largest deviation from normality appears to be with the smaller values; they are farthest from the horizontal line that goes through 0.

## Statistical Tests of Normality

The statistical tests for normality that are provided in SPSS for Windows (v. 22–23) are the Shapiro-Wilks and Kolmogorov-Smirnov (K-S) Lilliefors statistics. These can be obtained by selecting the *Analyze . . . Descriptive Statistics . . . Explore* commands from the menu, clicking on *Plots* and selecting *Normality plots with tests*. The objectives of these nonparametric goodness-of-fit tests are to compare the obtained distribution with a theoretically normal distribution of the same mean and standard deviation and to determine whether the deviations from normality are sufficiently large to conclude that the distribution under investigation is not normal. The null hypothesis is that the data are normally distributed; the alternative hypothesis is that the data are not normally distributed. The null hypothesis will be rejected if the obtained significance level is less than our stated level of alpha (e.g., $\alpha = .05$).

Both the Shapiro-Wilks and K-S Lilliefors statistics are extremely sensitive to departures from normality. It is strongly recommended, therefore, that the researcher supplement these statistical tests with the previously discussed methods for examining data for departures from normality (e.g., visually examining the data and assessing skewness and kurtosis).

The computer printout generated from SPSS for Windows for the Shapiro-Wilks and K-S Lilliefors statistics is presented in Table 3.2. For the preintervention fatigue variable, we have obtained similar results. Both tests indicate that the distribution is not normal (significance < .001 is less than $\alpha = .05$) (Table 3.2) ⑬. This is not always the case, however. Sometimes you will find that the two statistics disagree. Conover (1999) suggests that the Shapiro-Wilks test for normality may be more powerful than the K-S Lilliefors statistic in that it may be more likely to correctly reject the null hypothesis of normality.

**Figure 3.6**    Normal probability plots of preintervention fatigue scores (n = 20).
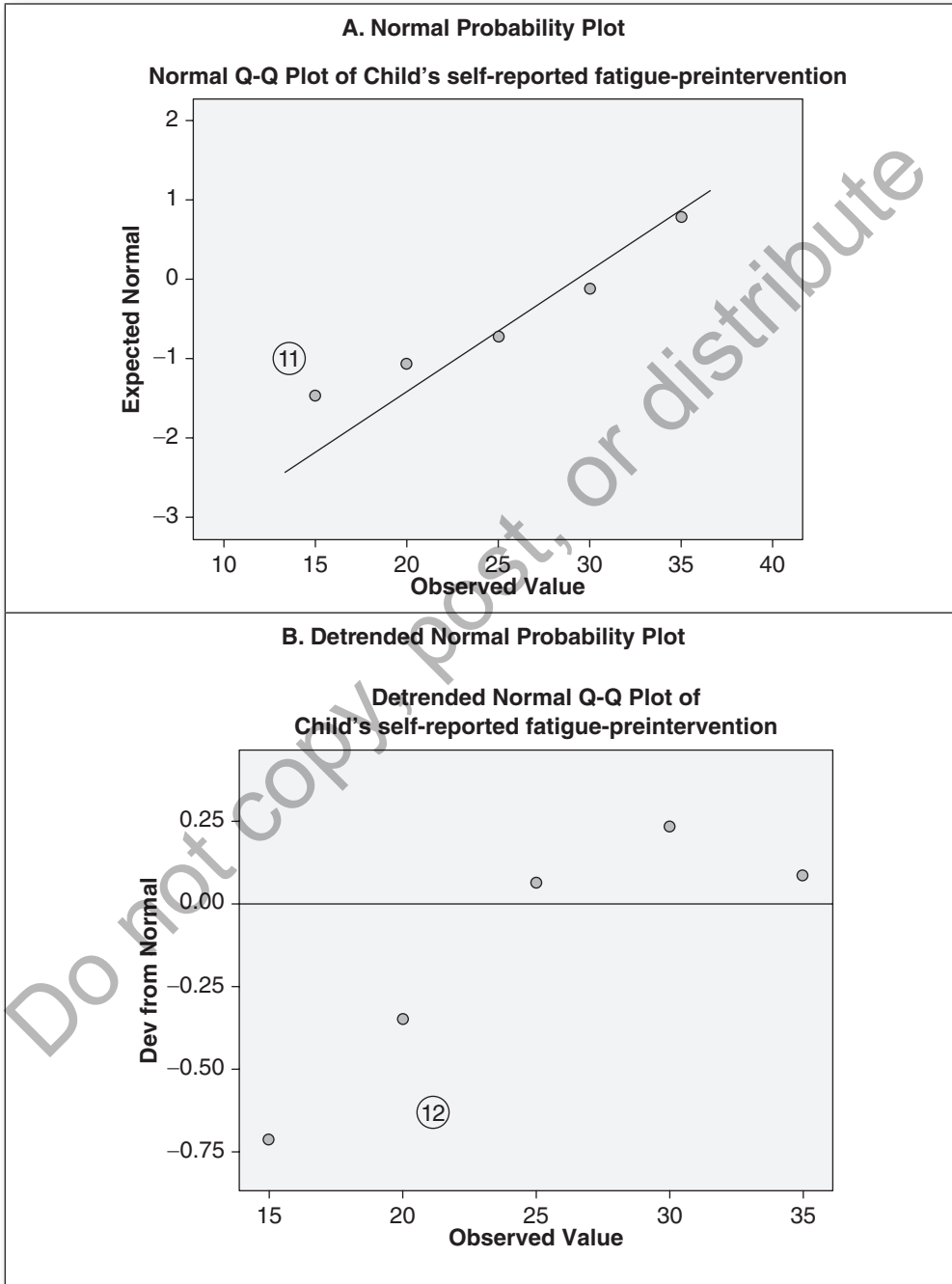


**A. Normal Probability Plot**

**Normal Q-Q Plot of Child's self-reported fatigue-preintervention**

**B. Detrended Normal Probability Plot**

**Detrended Normal Q-Q Plot of Child's self-reported fatigue-preintervention**

| Table 3.2 | Statistical Tests for Normality of the Preintervention Fatigue Variable |

**Tests of Normality**

| | Kolmogorov-Smirnov[a] | | | Shapiro-Wilk | | |
|---|---|---|---|---|---|---|
| | Statistic | df | Sig. | Statistic | df | Sig. |
| Child's self-reported fatigue-preintervention | .2.46 | 20 | ⑬ .003 | .812 | 20 | ⑬ .001 |

Reprints Courtesy of International Business Machines Corporation, © International Business Machines Corporation

[a]Lilliefors significance correction.

Our determination of whether to accept or reject the preintervention fatigue distribution as normal should be based on all contributing factors: the level of measurement of the data, its visual representation, the similarity of the measures of central tendency, skewness and kurtosis, the statistics, and the sample size. Based on this evidence, we would most likely conclude that the data for preintervention fatigue are *not* normally distributed. This conclusion is based on the observation that although the data might be considered interval level of measurement, the visual representations suggest nonnormality; the mean, median, and mode are not similar; there is some skewness; the Shapiro-Wilks and K-S Lilliefors statistics support rejection of the null hypothesis of normality; and we had a sample size of only 20. This determination would suggest that we would seriously need to consider using nonparametric statistics when analyzing this variable.

## Examining Distributions of the Dependent Variable by Subgroups

For many parametric tests, it is expected that the distribution of the dependent variable be normally distributed not only as a whole but also when broken down into subgroups of a particular independent variable of interest. Table 3.3 presents the syntax commands and a breakdown of the preintervention fatigue scores of the children by staff-initiated intervention and usual-care groups using the hospitalized children with cancer-20 cases.sav. These printouts were generated in SPSS for Windows by highlighting the *Analyze . . . Descriptive Statistics . . . Explore* commands (see Figure 3.2) and placing the dependent variable, *Intensity_Fatigue_preintervention,* in the *Dependent List* and the independent variable, *Group,* in the *Factor List.*

The resulting descriptive statistics (Table 3.3) and histograms (Figure 3.7) indicate that the staff-initiated intervention and usual-care groups have similar means and distributions. This suggests that we may have been successful in creating similar groups through random assignment—at least with regard to preintervention fatigue. The skewness statistics for the intervention group (skewness/standard error for skewness = −1.085/.687 = −1.579) and the usual-care group (−1.338/.687 = −1.95) also indicate that the variable's skewness for both groups is within an acceptable range (±1.96)

**Table 3.3**   Computer-Generated Printout of Pretreatment Fatigue by Group (Usual Care, Staff-Initiated Intervention) (SPSS for Windows, v.22–23)

```
EXAMINE VARIABLES=Intensity_fatigue_preintervention BY group
    /PLOT BOXPLOT HISTOGRAM NPPLOT
    /COMPARE GROUPS
    /STATISTICS DESCRIPTIVES
    /CINTERVAL 95
    /MISSING LISTWISE
    /NOTOTAL.
```

**Descriptives**

| | staff-initiated intervention vs. usual care | | | Statistic | Std. Error |
|---|---|---|---|---|---|
| Child's self-reported fatigue-preinteivention | usual care group | Mean | | 29.5000 | 2.03443 |
| | | 95% Confidence Interval for Mean | Lower Bound | 24.8978 | |
| | | | Upper Bound | 34.1022 | |
| | | 5% Trimmed Mean | | 30.0000 | |
| | | Median | | 30.0000 | |
| | | Variance | | 41.389 | |
| | | Std. Deviation | | 6.43342 | |
| | | Minimum | | 15.00 | |
| | | Maximum | | 35.00 | |
| | | Range | | 20.00 | |
| | | Interquartile Range | | 10.00 | |
| | | Skewness | | -1.338 | .687 |
| | | Kurtosis | | 1.864 | 1.334 |
| | staff-initiated intervention | Mean | | 29.0000 | 2.21108 |
| | | 95% Confidence Interval for Mean | Lower Bound | 23.9982 | |
| | | | Upper Bound | 34.0018 | |
| | | 5% Trimmed Mean | | 29.4444 | |
| | | Median | | 30.0000 | |
| | | Variance | | 48.889 | |
| | | Std. Deviation | | 6.99206 | |
| | | Minimum | | 15.00 | |
| | | Maximum | | 35.00 | |
| | | Range | | 20.00 | |
| | | Interquartile Range | | 11.25 | |
| | | Skewness | | -1.085 | .687 |
| | | Kurtosis | | .265 | 1.334 |

(14)
−1.338/0.687 = −1.95
−1.085/0.687 = −1.579

**Tests Df Normality**

| | staff-initiated inteivention vs. usual care | Kolmogorov-Smirnov[a] | | | Shapiro-Wilk | | |
|---|---|---|---|---|---|---|---|
| | | Statistic | df | Sig. | Statistic | df | Sig. |
| Child's self-reported fatigue-preintervention | usual care group | .231 | 10 | (16) .139 | .824 | 10 | (15) .028 |
| | staff.initiated intervention | .257 | 10 | .060 | .835 | 10 | .038 |

a. Lilliefors Significance Correction

Reprints Courtesy of International Business Machines Corporation, © International Business Machines Corporation

**Figure 3.7**   Boxplots of preintervention fatigue for the staff-initiated and usual-care group generated in SPSS for Windows (v. 22–23).
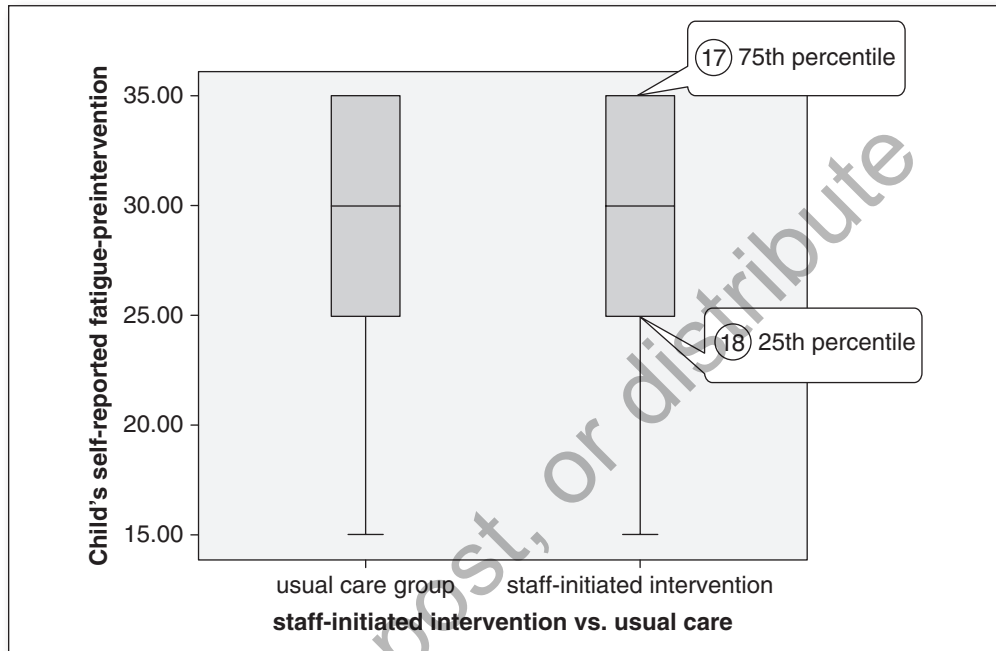


Reprints Courtesy of International Business Machines Corporation, © International Business Machines Corporation

(Table 3.3, ⑭). Given the small sample size for both groups ($n = 10$), however, as well as the shape of the histograms for both groups, nonparametric tests most likely would be used with these data. This conclusion is further supported by the significant Shapiro-Wilks tests for both groups, .028 and .038, which are less than $\alpha = .05$ ⑮.

Notice, however that the Kolmogorov-Smirnov $p$ values for the two groups are .139 and .06 ⑯, both of which are >.05. This test suggests that we should retain the null hypothesis, which states that the data are normally distributed. What should we do with this conflicting advice? Again, we need to return to the plots of the data (Figure 3.7) to determine for ourselves which of these two statistics we should believe. The results presented in Figure 3.7 suggest that both of the distributions for the usual-care and intervention groups appear to be negatively skewed. The conclusion, therefore, would be that, indeed, we do have skewed distributions for both groups.

## DEALING WITH OUTLIERS

One of the disadvantages of the mean as a measure of central tendency is its sensitivity to outliers. Because outliers are extreme data points that are very much different from the rest of the data, they tend to pull the value of the mean in their direction. This can

result in serious distortion of results. The median, on the other hand, is not at all influenced by atypical data points because the median assesses ranks, not actual values. The presence of outliers, therefore, requires a careful assessment of their influences both on the mean and on the variable's distribution. Outliers also provide information about the types of cases that may not fit a particular hypothesized model.

There are two types of outliers: univariate and multivariate. Univariate outliers are those cases that possess extreme values on a single variable (e.g., a child who has an extreme fatigue score). Multivariate outliers are cases with unusual combinations of scores on two or more variables. For example, a person may be of an acceptable age (e.g., 16 years old) and another person could have a reasonable number of children (e.g., four), but a 16-year-old who has four children would most likely appear as a multivariate outlier.

## Assessing Univariate Outliers Using the Boxplot

Boxplots (Figure 3.7) are very useful for identifying cases that are univariate outliers. They also provide a snapshot summary of the descriptive statistics for the distribution. On request, SPSS for Windows plots the smallest and largest values of the data set, the median (the horizontal bar inside the box), the 25th percentile (the lower boundary of the box), and the 75th percentile (the upper boundary), and it presents values that lie far outside this range. The interquartile range makes up the box presented in this plot. This is where 50% of the cases are located. The boxplot for the normal distribution in Figure 3.5 A illustrates a distribution that is symmetrical, with equal tails, and a median that lies halfway between the upper and lower boundaries of the box.

Two types of univariate outliers are presented in the boxplots for SPSS for Windows. Any value that is more than three box-lengths (i.e., $3[P_{75} - P_{25}]$) from the upper or lower boundary of the box is designated on the plot with a "*" and is referred to as an *extreme value*. Each value that is between 1.5 (i.e., $1.5[P_{75} - P_{25}]$) and 3 box-lengths from the upper or lower boundary of the box is identified with an "O" and is called an *outlier*. The outliers and extreme values are also identified either by their case number (the default option) or by specifying a case label (e.g., the variable *id*). This information is useful for tracking down and correcting possible errors in data entry. The largest and smallest observed values that are not outliers are presented by lines drawn from the ends of the box to these values.

In general, boxplots are useful for comparing the distribution of a continuous variable for two or more subgroups in a sample. For example, Figure 3.5, in panels B to D, presents the boxplots for a positively skewed, a negatively skewed, and a bimodal distribution. The boxplots for the positively and negatively skewed distributions indicate that the distributions are asymmetrical, having a long tail in one direction. The median in each case is no longer in the middle of the box but rather lies closer to the bottom or top of the box, depending on the type of skew. Extreme values (*) and outliers (O) can also be found lying beyond the longer tail. It is interesting that the boxplot for a bimodal distribution (Figure 3.5D) is not very helpful in revealing the shape of the distribution. Although the box for this distribution is very large compared to the tails and there are no outliers, its bimodal shape has become hidden.

Boxplots are especially useful for comparing two distributions. For example, the boxplots for the preintervention fatigue scores for the staff-initiated intervention and usual-care groups are presented in Figure 3.7. These boxplots confirm our suspicion, based on visual inspection, that the preintervention fatigue data are negatively skewed for both groups: There is only one tail presented, directed toward the lower end of the values. Had the data been more normally distributed, two tails of equal length would have been presented, and the boxplots would have been similar to that in Figure 3.5A.

The lack of an upper tail for the preintervention anxiety scores in Figure 3.7 is understandable because there is a restricted range for this variable (14–35). For the staff-initiated intervention group, for example, the 75th percentile for this distribution is identified in the graph as the value of 35 (17) and the 25th percentile as the value 25 (18). Because 3 box-lengths is equal to 30 (3[$P_{75} - P_{25}$] = 3 [35 − 25] = 3 [10] = 30 and 1.5 box-lengths is equal to 15 (1.5[$P_{75} - P_{25}$] = 1.5[35 − 25] = 1.5[10] = 15), the extreme values (*) for this example would be those values that are either 65 or larger (35 + 30=65) or -5 or smaller (25 − 30 = −5). Outliers (O) would be 1.5 box-lengths above and below the upper and lower boundaries of the box, or the values of 50 (35 + 15 = 50) and 10 (25 − 15 = 10) respectively.

No children reported scores of less than 14 or higher than 35, so there were no outliers. Because there were no extreme values or minor outliers for this distribution, there are no "*" or "O" symbols in the computer printout. The conclusion to be drawn, therefore, is that the distribution of these data for both groups is relatively compact, of low range, and not normal.

## Assessing Multivariate Outliers

Although the boxplot provides useful information about univariate outliers, it does not tell us anything about cases that have unusual patterns of scores with respect to two or more variables. These multivariate outliers can be screened by computer using techniques made available within SPSS using its regression analyses. Because the focus of this text is on nonparametric statistics, we will not examine these issues here. For the interested reader, these techniques (e.g., examining linear relationships, use of the Mahalanobis distance, and approaches to the analyses of residuals) are described in great detail and clarity by Hair, Black, Babin, Anderson, and Tatham (2010); J. Stevens (2009); and Tabachnick and Fidell (2013) in their excellent textbooks on multivariate statistical analysis.

## What to Do About Outliers

Researchers appear to have mixed feelings about outliers and what to do about them. Some researchers view outliers as nuisance cases, ones that do not fit expectations. Others suggest that the outliers in a study are the cases that should be examined most closely. Kruskal (1988), for example, argues that "miracles are the extreme outliers of nonscientific life. . . . It is widely argued of outliers that investigation of the mechanism for outlying may be far more important than the original study that led to the outlier" (p. 929).

A critical task for the researcher is to determine why outliers exist in the first place. Are they a result of errors of coding or measurement, or are they legitimate cases that possess unique characteristics with respect to one or more variables? Different approaches to remedying problematic outliers and reducing their influence have been suggested, depending on the etiology of the outlier's presence (Hair et al., 2010; Johnson, 1985; Pedhazur & Schmelkin, 1991; Tabachnick & Fidell, 2013). Such techniques include eliminating the case altogether, reweighting or recoding the outlier to reduce its influence, and transforming the variable to create a more nearly normal distribution. It may also be useful to analyze the data both with and without the extreme data points to determine the extent of the outliers' influence.

An enormous advantage of nonparametric rank-order statistics is that the ranking of data that occur with these statistics serves to reduce the influence of outliers because the data being analyzed are ranks, not actual scores. There is no "quick fix" to the problem of outliers, and careful attention must be paid to the consequences of a particular remedy. These decisions must also be duly reported in the data analyses.

## DATA TRANSFORMATION CONSIDERATIONS

When a particular distribution of a variable does not meet the normality assumption, it is possible to transform the values of that variable to create a new variable that has a more nearly normal distribution. Although this process appears easily accomplished, it does have serious problems, particularly with regard to both finding an adequate transformation index that will produce a more nearly normal distribution and interpreting the results of such a transformation. Figure 3.3 presents several common forms of nonnormal distributions and some suggested transformations that might help to create a more nearly normal distribution for the transformed variable. Hair et al. (2010) suggest that for flat (platykurtic) distributions (Figure 3.3E), the most common transformation is the inverse $(1/x)$. A variable that is positively skewed (Figure 3.3B) might benefit from a log transformation $(\log(x))$, whereas one that is negatively skewed (Figure 3.3C) might be altered with a square root transformation. Leptokurtic distributions (Figure 3.3D) do not appear to have clearly defined transformations available in the research literature. Hair et al. (2010) also indicate that to achieve a noticeable effect from a transformation, the ratio of a variable's mean to its standard deviation should be less than 4.0 (i.e., mean/standard deviation < 4.0).

The goal of transforming data is to obtain a new distribution that is nearly normal in shape, with few outliers, and with skewness and kurtosis values near zero. It is important, therefore, that the researcher closely examine the distribution of the resulting transformation to ascertain if this goal has been achieved. Next, a careful interpretation of the resulting transformation needs to be made. Remember that a transformed variable no longer carries the original interpretation; the square root of preintervention fatigue is *not* the same as preintervention fatigue. Interpreting the meaning of a transformed variable is one of the most challenging tasks for the researcher.

In an attempt to obtain a more nearly normal distribution, the preintervention fatigue variable was transformed using two suggested transformations for negatively skewed distributions (Figure 3.3C). First we reflected the original variable such that the scores were reversed (i.e., *new score = (largest old score +1) – old score*), and then we took the square root and log of this newly created variable. We are using the "reflect" because our data are negatively skewed. The reflect allows us to reverse code the old variable and then take a square root or a log of the newly created variable. We need to be extremely careful, however, in our interpretation of this newly created variable since the interpretation of the direction of the scoring is now opposite of what it was before. If, for the untransformed variable, higher scores meant *greater* fatigue, then higher scores on this transformed variable will mean *lower* fatigue.

Transformations of variables can be undertaken easily in SPSS for Windows through its *Transform . . . Compute Variable* command (Figure 3.8). Using the data set, hospitalized children with cancer-20 cases.sav, two new target variables, *reflect_sqrt_fatigue_t1* and *reflect_log_fatigue_t1*, were obtained by indicating that they represent the reflect of the square root (and log) of the old variable, *Intensity_fatigue_preintervention*.

Figure 3.9 compares the newly formed reflect of the square root and log transformations with the original preintervention fatigue distribution. If the goal of data transformation is to obtain a nearly normal distribution with few outliers and with values of skewness and kurtosis near zero, it is apparent that while these transformations succeeded in lowering the skewness coefficients (Figure 3.9) to below the ±1.96 range, the shape of the resulting distributions is not normal. This failure to produce a more normal distribution may be a result of the small sample size ($n = 20$) and limited scale values (14–35). It also suggests that nonparametric statistics, which rely predominantly on the ranking of data, may be the approach of choice.

## EXAMINING HOMOGENEITY OF VARIANCE

Another important assumption of parametric tests that compare differences between two or more groups is that the variances among the subgroups must be similar; that is, there is homogeneity of variance. A general rule of thumb is that the variance of one group should not be more than twice that of another. This assumption is especially important when groups of unequal size are being compared (Tabachnick & Fidell, 2013).

Several tests of homogeneity of variance are available in SPSS. These include Box's *M* and the Levene test. The null hypothesis for all tests of homogeneity is that the variances among the groups are equal, whereas the alternative hypothesis states that the variances are unequal. The null hypothesis will be rejected if the obtained level of significance is less than the preset level of alpha (e.g., $\alpha = .05$).

The descriptive statistics presented for the preintervention fatigue variable in Table 3.3 indicate that the variance for the usual-care group was 41.39 compared to 48.89 for the intervention group. Because one variance is less than twice the other, it would appear that the homogeneity of variance assumption for preintervention fatigue has been met. The resulting Levene test generated from the *Analyze . . . Compare Means . . . Independent Samples T-test* command indicates that we would indeed fail to reject the null hypothesis
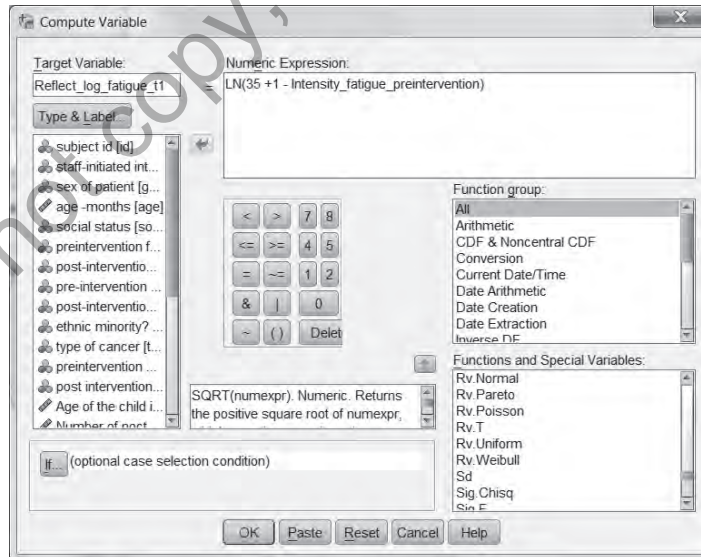
**Figure 3.8** SPSS for Windows commands for transforming the negatively skewed fatigue variable.

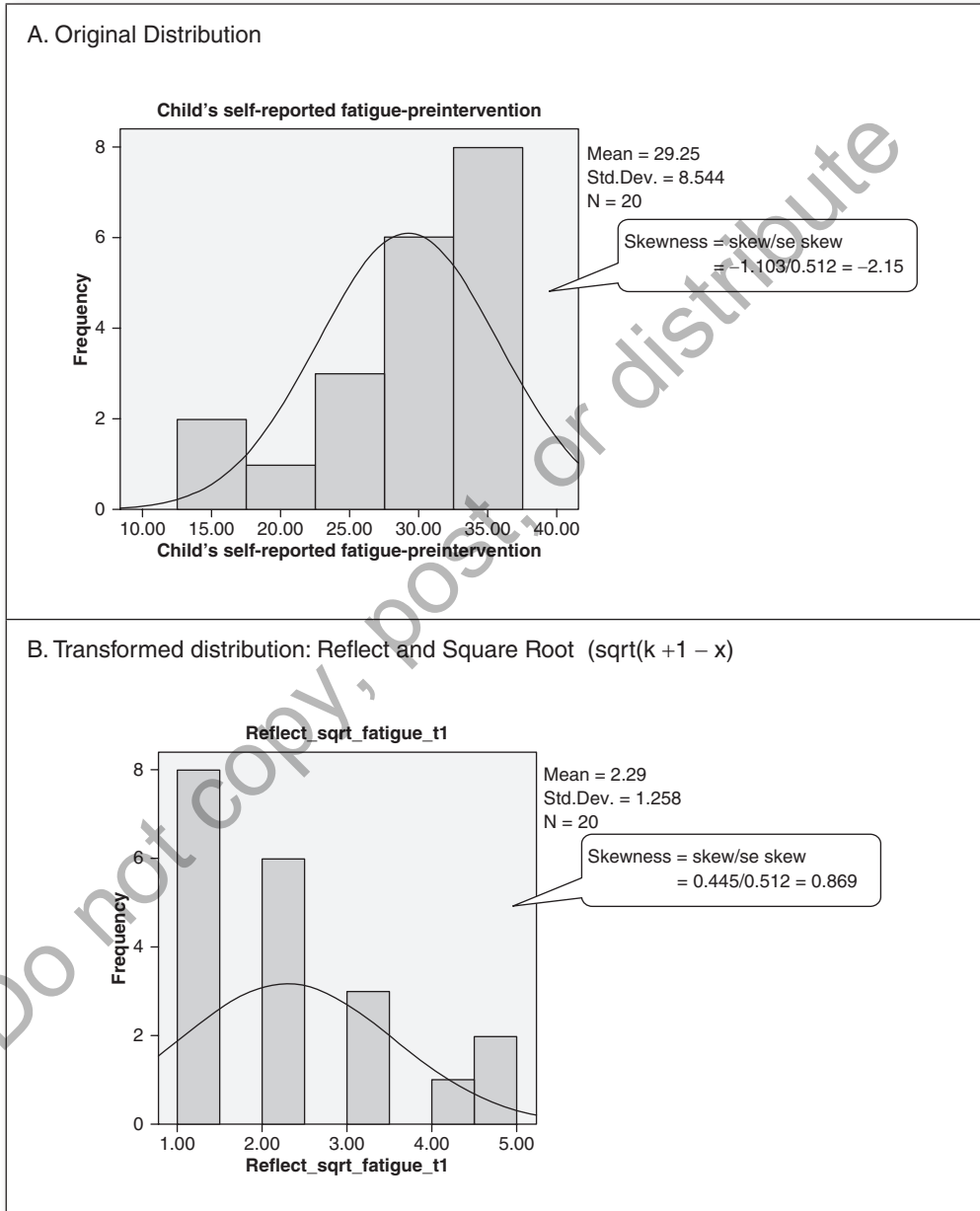A. Square Root of the Reflected Fatigue Variable
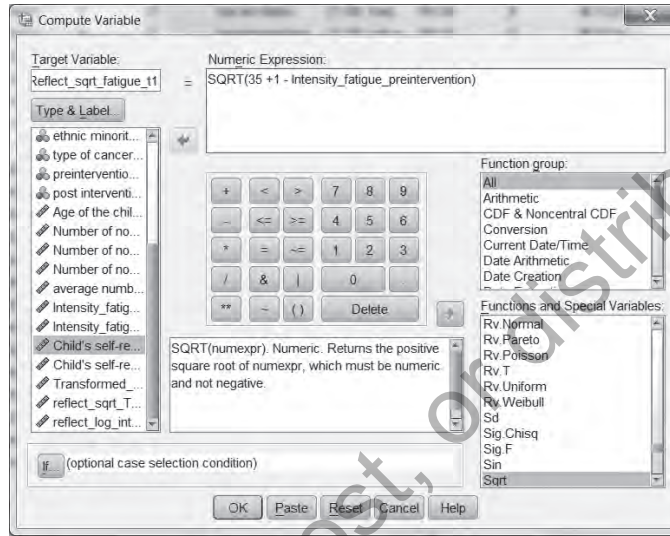
B. Log of the Reflected Fatigue Variable

Reprints Courtesy of International Business Machines Corporation, © International Business Machines Corporation
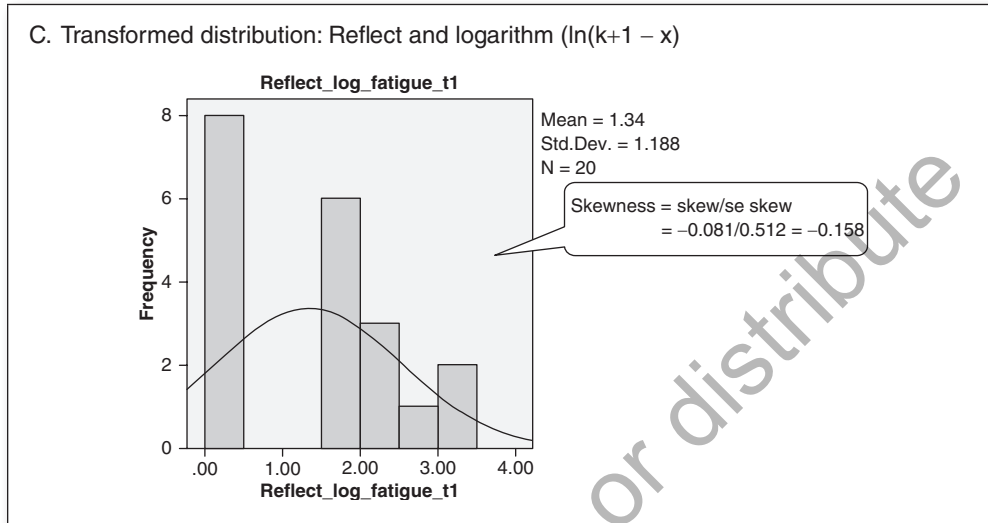
**Figure 3.9**   Results of reflect transformations of the preintervention fatigue variable.



A. Original Distribution

**Child's self-reported fatigue-preintervention**

Mean = 29.25
Std.Dev. = 8.544
N = 20

Skewness = skew/se skew
= −1.103/0.512 = −2.15

B. Transformed distribution: Reflect and Square Root  (sqrt(k +1 − x)

**Reflect_sqrt_fatigue_t1**

Mean = 2.29
Std.Dev. = 1.258
N = 20

Skewness = skew/se skew
= 0.445/0.512 = 0.869

*(Continued)*

| Figure 3.9 | (Continued) |
|---|---|

C. Transformed distribution: Reflect and logarithm (ln(k+1 − x)



Reflect_log_fatigue_t1

Mean = 1.34
Std.Dev. = 1.188
N = 20

Skewness = skew/se skew
= −0.081/0.512 = −0.158

of equal variances because the significance level (.708) is considerably greater than our $\alpha$ = .05. We should be pleased with this "failure" because we can conclude that the variances between the groups are equal.

## EVALUATING SAMPLE SIZES

Two additional issues that are important determinants of which statistic—parametric or nonparametric—is most appropriate to use in an analysis are the absolute size of the sample being used and whether there are equal numbers of subjects in the subgroups being analyzed.

### Minimum Sample Size Requirements

Determining an appropriate sample size for a study is a challenging task and must be evaluated with each planned statistical test. Sample size consideration is also directly related to statistical power, or the researcher's ability to correctly reject a null hypothesis. All other factors being equal (e.g., similar effect size and alpha level), the larger the sample size, the more likely it is that the researcher will be able to detect significant differences among groups. Smaller sample sizes reduce statistical power and increase the researcher's chance of making a Type II error or incorrectly concluding that the null hypothesis is true.

There is no definitive guideline as to how large a *large* sample is or what size sample is *too small* for use with a particular parametric test. The process of determining "how large is large enough?" is inextricably linked to issues of statistical power, the selected level of alpha, and the projected size of the expected differences among the groups (i.e., the "effect size").

Determining an adequate sample size also needs to be focused on the requirements of specific statistical tests (e.g., multiple regression, *t* tests, and ANOVA). An "adequate" sample size also depends in part on the normality of the distribution of the dependent variable of interest. It should be noted, though, that whatever the form of the original distribution, the central limit theorem has shown that as the sample size increases, the sampling distribution of the *mean* approaches normality. Because the mean plays a major role in most parametric statistics, this characteristic of the shape of the distribution of the mean is especially important.

In their classic text on nonparametric statistics, Siegel and Castellan (1988) argued that "if the sample size is very small, there may be no alternative to using a nonparametric statistical test unless the nature of the population distribution is known exactly" (p. 35). Unfortunately, the authors did not state specifically what they meant by a "very small" sample size. Other authors have discussed implications of using parametric statistical tests when sample sizes are *exceedingly* or *extremely* small (Neter et al., 1993) but, again, these limits are not specifically defined. Where does that leave us, as researchers, when we are faced with a total sample size of 20 and no likelihood of increasing the size? Do we use parametric or nonparametric tests to analyze our data?

The answer to this question is that "It depends." That is, the choice of a parametric or nonparametric test depends on the available sample, the variables' level of measurement, the shape of the distributions of the variables of interest, and the researcher's knowledge of the variables' distribution in the population. In addition, the researcher needs to consider the number of subgroups to be examined in the analyses. For example, dividing our sample of 20 hospitalized children into four subgroups depending on their type of cancer (e.g., solid tumor, acute myeloid leukemia, lymphoma, and sarcoma) will result in considerably smaller sample sizes (range: 2–8) to analyze than will merely breaking the group down by experimental and control groups (*n* = 10). Moreover, parametric analyses of the interaction of these two variables (group by type of minor surgery) would be rendered impossible given the resulting small and unequal sample sizes (*n* = 2 or 3 per group) within each subcategory of group by type of surgery interaction.

Assuming that we have sufficient funds to carry out this pilot study, should we consider increasing our sample size? Again, sample size determination is dependent on a number of factors (e.g., alpha, desired power, effect size, and planned statistical tests). Luckily, there are several computer programs available to help the researcher to undertake a statistical power analysis to determine the most appropriate sample size for most parametric and nonparametric statistical tests. These programs include G*Power 3 (Faul, Erdfelder, Lang, & Buchner, 2007), PASS 12 (Hintze, 2013), and IBM® SPSS® SamplePower (v. 3.0.1) (IBM® SPSS®, 2015).

## Equal Numbers of Subjects Within Subgroups

As a general rule of thumb, it is highly desirable to have equal numbers of subjects within subgroups. Although this goal is attainable in experimental and quasi-experimental designs, it is often violated in descriptive studies because of natural selection. It is, therefore, a topic that cannot be avoided in health care research.

Unequal cell sizes have a differential impact on the results of a research study depending on the statistics selected. For example, the problems raised by unequal

numbers of subjects are not disastrous when using *t* tests or simple one-way ANOVAs, particularly if the condition of homogeneity of variance is met (Tabachnick & Fidell, 2013). Serious problems regarding unequal cell sizes with or without homogeneity of variance arise, however, in factorial designs, such as two-way ANOVA, analysis of covariance (ANCOVA), multivariate analysis of variance (MANOVA), repeated-measures ANOVA, and multiple regression. These problems are generated because the hypotheses that test the main effects and interaction terms are no longer orthogonal (i.e., independent) when there are unequal cell sizes. Tabachnick and Fidell (2013) offer an in-depth discussion of both the problem of unequal cell sizes and possible solutions in their text on multivariate statistics. The problem of confounding of main effects for multiple factors when there are unequal cell sizes is present in some nonparametric analyses as well.

## REPORTING TESTING ASSUMPTIONS AND VIOLATIONS IN A RESEARCH REPORT

The reporting of assessment of assumptions and their violations does not appear to be a common practice in health care research. Pett and Sehy (1996) found that less than 20% of the 238 randomly selected nursing research articles that they reviewed included even a brief discussion (i.e., a sentence or two) of the assumptions underlying the use of the statistical tests reported. Discussion of the formal testing of the assumptions underlying the statistical tests, management of violations of these assumptions, and handling of extreme data was also uncommon (15.5%, 14.7%, and 3.4%, respectively). Similar findings were reported by Gaither and Glorfeld (1983) in organizational behavior. These authors offer the hope that researchers do indeed consider the assumptions underlying their choice of statistical tests carefully but either choose not to report their findings or are discouraged to do so by reviewers or editors of journals.

The reporting of tests of assumptions and handling of violations does not have to be an arduous undertaking. A mere sentence or two would suffice. Robichaud-Ekstrand (1991), for example, did this very nicely in her report of responses to shower versus sink baths for 30 patients with myocardial infarction:

> Because the distribution of the overall subjective scores was skewed, a Friedman test ($p \le .05$) and Wilcoxon post hoc pairwise tests ($p \le .02$) were used. Pearson [product moment] correlations were used to compare the skewed scores of RPE [ratings of perceived exertion] and the overall subjective scores with the normally distributed HR [heart rate] and BP [blood pressure] scores. (p. 378)

## Summary

This chapter has attempted to provide the reader with some guidelines for determining which statistics, parametric or nonparametric, are best suited for the data at hand. Box 3.1 presents a summary checklist of the procedures that the researcher can use to assess the characteristics of the data being analyzed. These steps include determining the level of

measurement of the variables of interest, evaluating their distributions, assessing homogeneity of variances, considering sample sizes, determining the statistics that are best suited for these data, and duly reporting the results of this investigation. It is rare that data collected from the "real world" are perfectly suited to the requirements of a particular parametric or nonparametric test. Remember, however, that the bottom line to the choice of most parametric statistics is the important question: *Do the mean and standard deviation truly represent your data?*

---

### BOX 3.1   Checklist for Assessing the Characteristics of Data

1. Determine the levels of measurement of the variables of interest.

2. Evaluate the distribution of these variables:

   - Compare the measures of central tendency for each variable
   - Determine skewness and kurtosis
   - Visually assess the distributions
   - Examine the probability plots
   - Transform the variables if necessary
   - Examine and interpret the results of the transformation

3. Check for homogeneity of variance.

4. Assess the total sample size and size of the subgroups.

5. Determine which statistic, parametric or nonparametric, is best suited for these data.

6. Report the process of evaluation and decisions in the data analysis section.

---

### TEST YOUR KNOWLEDGE

Here are some questions that should help you recall the key points of this chapter. If you find that you cannot recall the answers, go back to the text and reread the appropriate sections.

1. What are the characteristics of the four levels of measurement?

2. What are the levels of measurement of the following variables?

   a. Time to wakefulness following anesthesia

   b. Weight of a child in pounds

   c. Scores on a single-item pain scale (0 = *little or no pain* to 10 = *severe pain*)

   d. IQ of your significant other

   e. Staging of breast cancer (Stages I through IV)

   f. A list of different specialties in your profession

*(Continued)*

(Continued)

     g. Level of depression as measured by a 30-item (0 = no, 1 = yes) depression scale (total score: 0–30)

     h. Ethnicity of the patient

3. What is the meaning of skewness and kurtosis, and how would you evaluate these conditions?

4. What is the difference between a univariate and multivariate outlier? How would you assess these conditions?

5. What kind(s) of transformations might I consider if I had the following distribution for a variable:

     a. Positively skewed distribution

     b. Negatively skewed distribution

     c. Mesokurtic distribution

     d. J-shaped distribution

## COMPUTER EXERCISES

Here is a computer exercise that should help you to put into practice that which has been discussed in this chapter.

1. Using the data set provided to you at study.sagepub.com/pett2e (use hospitalized children with cancer-20 cases.sav or hospitalized children with cancer-20 cases.xlsx), please undertake the following assessments:

     a. Using SPSS for Windows (or another suitable statistical program), generate descriptive statistics including skewness and kurtosis coefficients, boxplots, and histograms for the following variables: age of the child in years, preintervention fatigue, and postintervention fatigue.

     b. In a sentence or two, describe the distributions of these three variables. Are there any outliers that you are concerned about?

     c. Which of these variables would you consider transforming? If so, what type of transformation would you choose?

     d. Undertake the transformation(s) of these variables. Rerun the descriptive statistics on these variables. How, if at all, did the transformation alter the shape, skewness, and kurtosis of your variables?

Visit **study.sagepub.com/pett2e** to access SAS output, SPSS datasets, SAS datasets, and SAS examples.