evaluate the potential generalizability of research results based on convenience samples. It is possible to imagine a hypothetical population—that is, a larger group of people that is similar in many ways to the participants who were included in the convenience sample—and to make cautious inferences about this hypothetical population based on the responses of the sample. Campbell suggested that researchers evaluate the degree of similarity between a sample and hypothetical populations of interest and limit generalizations to hypothetical populations that are similar to the sample of participants actually included in the study. If the convenience sample consists of 50 Corinth College students who are between the ages of 18 and 22 and mostly of Northern European family background, it might be reasonable to argue (cautiously, of course) that the results of this study potentially apply to a hypothetical broader population of 18- to 22-year-old U.S. college students who come from similar ethnic or cultural backgrounds. This hypothetical population—all U.S. college students between 18 and 22 years from a Northern European family background—has a composition fairly similar to the composition of the convenience sample. It would be questionable to generalize about response to caffeine for populations that have drastically different characteristics from the members of the sample (such as persons who are older than age 50 or who have health problems that members of the convenience sample do not have).

Generalization of results beyond a sample to make inferences about a broader population is always risky, so researchers should be cautious in making generalizations. An example involving research on drugs highlights the potential problems that can arise when researchers are too quick to assume that results from convenience samples provide accurate information about the effects of a treatment on a broader population. For example, suppose that a researcher conducts a series of studies to evaluate the effects of a new antidepressant drug on depression. Suppose that the participants are a convenience sample of depressed young adults between the ages of 18 and 22. If the researcher uses appropriate experimental designs and finds that the new drug significantly reduces depression in these studies, the researcher might tentatively say that this drug may be effective for other depressed young adults in this age range. It could be misleading, however, to generalize the results of the study to children or to older adults. A drug that appears to be safe and effective for a convenience sample of young adults might not be safe or effective in patients who are younger or older.

To summarize, when a study uses data from a convenience sample, the researcher should clearly state that the nature of the sample limits the potential generalizability of the results. Of course, inferences about hypothetical or real populations based on data from a single study are never conclusive, even when random selection procedures are used to obtain the sample. An individual study may yield incorrect or misleading results for many reasons. Replication across many samples and studies is required before researchers can begin to feel confident about their conclusions.

### 1.6 Levels of Measurement and Types of Variables

A controversial issue introduced early in statistics courses involves types of measurement for variables. Many introductory textbooks list the classic levels of measurement defined by Stevens (1946): nominal, ordinal, interval, and ratio (see Table 1.1 for a summary...
and Note 3 for a more detailed review of these levels of measurement). Strict adherents to the Stevens theory of measurement argue that the level of measurement of a variable limits the set of logical and arithmetic operations that can appropriately be applied to scores. That, in turn, limits the choice of statistics. For example, if scores are nominal or categorical level of measurement, then according to Stevens, the only things we can legitimately do with the scores are count how many persons belong to each group (and compute proportions or percentages of persons in each group); we can also note whether two persons have equal or unequal scores. It would be nonsense to add up scores for a nominal variable such as eye color (coded 1 = Blue, 2 = Green, 3 = Brown, 4 = Hazel, 5 = Other) and calculate a “mean eye color” based on a sum of these scores.

In recent years, many statisticians have argued for a much less strict application of level of measurement requirements. In practice, there are many common types of variables (such as 5-point rating scales for attitude and personality measurement, probably fall short of satisfying the requirement that equal differences between scores represent exactly equal changes in the amount of the underlying characteristics being measured. However, most authors (such as Harris, 2001) argue that application of parametric statistics to scores that fall somewhat short of the requirements for interval level of measurement does not necessarily lead to problems.

Table 1.1 ♦ Levels of Measurement, Arithmetic Operations, and Types of Statistics

<table>
<thead>
<tr>
<th>Stevens’s Levels of Measurement</th>
<th>Logical and Arithmetic Operations That Can Be Applied (According to Stevens)</th>
<th>Traditional or Conservative Recommendation</th>
<th>Simpler Distinction Between Two Types of Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominal</td>
<td>=, ≠</td>
<td>Only nonparametric statistics</td>
<td>Categorical</td>
</tr>
<tr>
<td>Ordinal</td>
<td>=, ≠, &lt;, &gt;</td>
<td>Only nonparametric statistics</td>
<td>Quantitative</td>
</tr>
<tr>
<td>Intervalb</td>
<td>=, ≠, &lt;, &gt;, +, –</td>
<td>Parametric statistics</td>
<td>Quantitative</td>
</tr>
<tr>
<td>Ratio</td>
<td>=, ≠, &lt;, &gt;, +, –, ×, ÷</td>
<td>Parametric statistics</td>
<td>Quantitative</td>
</tr>
</tbody>
</table>

b. Many variables that are widely used in the social and behavioral sciences, such as 5-point rating scales for attitude and personality measurement, probably fall short of satisfying the requirement that equal differences between scores represent exactly equal changes in the amount of the underlying characteristics being measured. However, most authors (such as Harris, 2001) argue that application of parametric statistics to scores that fall somewhat short of the requirements for interval level of measurement does not necessarily lead to problems.
scores on many types of variables (such as 5-point rating scales) probably fall into a fuzzy region somewhere between the ordinal and interval levels of measurement. How crucial is it that scores meet the strict requirements for interval level of measurement?

Many statisticians have commented on this problem, noting that there are strong differences of opinion among researchers. Vogt (1999) noted that there is considerable controversy about the need for a true interval level of measurement as a condition for the use of statistics such as mean, variance, and Pearson $r$, stating that “as with constitutional law, there are in statistics strict and loose constructionists in the interpretation of adherence to assumptions” (p. 158). Although some statisticians adhere closely to Stevens's recommendations, many authors argue that it is not necessary to have data that satisfy the strict requirements for interval level of measurement in order to obtain interpretable and useful results for statistics such as mean and Pearson $r$.

Howell (1992) reviewed the arguments and concluded that the underlying level of measurement is not crucial in the choice of a statistic:

The validity of statements about the objects or events that we think we are measuring hinges primarily on our knowledge of those objects or events, not on the measurement scale. We do our best to ensure that our measures relate as closely as possible to what we want to measure, but our results are ultimately only the numbers we obtain and our faith in the relationship between those numbers and the underlying objects or events...the underlying measurement scale is not crucial in our choice of statistical techniques...a certain amount of common sense is required in interpreting the results of these statistical manipulations. (pp. 8–9)

Harris (2001) says,

I do not accept Stevens's position on the relationship between strength [level] of measurement and “permissible” statistical procedures...the most fundamental reason for [my] willingness to apply multivariate statistical techniques to such data, despite the warnings of Stevens and his associates, is the fact that the validity of statistical conclusions depends only on whether the numbers to which they are applied meet the distributional assumptions...used to derive them, and not on the scaling procedures used to obtain the numbers. (pp. 444–445)

Gaito (1980) reviewed these issues and concluded that “scale properties do not enter into any of the mathematical requirements” for various statistical procedures, such as ANOVA. Tabachnick and Fidell (2007) address this issue in their multivariate textbook: “The property of variables that is crucial to application of multivariate procedures is not type of measurement so much as the shape of the distribution” (p. 6). Zumbo and Zimmerman (1993) used computer simulations to demonstrate that varying the level of measurement for an underlying empirical structure (between ordinal and interval) did not lead to problems when several widely used statistics were applied.

Based on these arguments, it seems reasonable to apply statistics (such as the sample mean, Pearson $r$, and ANOVA) to scores that do not satisfy the strict requirements for interval level of measurement. (Some teachers and journal reviewers continue to prefer
the more conservative statistical practices advocated by Stevens; they may advise you to avoid the computation of means, variances, and Pearson correlations for data that aren't clearly interval/ratio level of measurement.)

When making decisions about the type of statistical analysis to apply, it is useful to make a simpler distinction between two types of variables: *categorical* versus *quantitative* (Jaccard & Becker, 2002). For a categorical or nominal variable, each number is merely a label for group membership. A **categorical variable** may represent naturally occurring groups or categories (the categorical variable gender can be coded 1 = Male, 2 = Female). Alternatively, a categorical variable can identify groups that receive different treatments in an experiment. In the hypothetical study described in this chapter, the categorical variable treatment can be coded 1 for participants who did not receive caffeine and 2 for participants who received 150 mg of caffeine. It is possible that the outcome variable for this imaginary study, anxiety, could also be a categorical or nominal variable; that is, a researcher could classify each participant as either 1 = Anxious or 0 = Not anxious, based on observations of behaviors such as speech rate or fidgeting.

**Quantitative variables** have scores that provide information about the magnitude of differences between participants in terms of the amount of some characteristic (such as anxiety in this example). The outcome variable, anxiety, can be measured in several different ways. An observer who does not know whether each person had caffeine could observe behaviors such as speech rate and fidgeting and make a judgment about each individual's anxiety level. An observer could rank order the participants in order of anxiety: 1 = Most anxious, 2 = Second most anxious, and so forth. (Note that ranking can be quite time-consuming if the total number of persons in the study is large.)

A more typical measurement method for this type of research situation would be self-report of anxiety, perhaps using a 5-point rating scale similar to the one below. Each participant would be asked to choose a number from 1 to 5 in response to a statement such as "I am very anxious."

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strongly disagree</td>
<td>Disagree</td>
<td>Neutral</td>
<td>Agree</td>
<td>Strongly agree</td>
</tr>
</tbody>
</table>

Conventionally, 5-point rating scales (where the five response alternatives correspond to "degrees of agreement" with a statement about attitude, belief, or behavior) are called **Likert scales**. However, rating-scale questions can have any number of response alternatives, and the response alternatives may have different labels—for example, reports of the frequency of a behavior. (See Chapter 19 for further discussion of self-report questions and response alternatives.)

What level of measurement does a 5-point rating scale similar to the one above provide? The answer is that we really don't know. Scores on 5-point rating scales probably do not have true equal interval measurement properties; we cannot demonstrate that the increase in the underlying amount of anxiety represented by a difference between 4 points and 3 points corresponds exactly to the increase in the amount of anxiety represented by the difference between 5 points and 4 points. It is not clear whether the "strongly disagree" response represents a true 0 point. When we try to decide whether a 5-point
scale is ordinal, interval, or ratio, we can give, at best, only an approximate answer. Five-point rating scales similar to the example above probably fall into a fuzzy category somewhere between the ordinal and interval levels of measurement. In practice, many researchers apply statistics such as means and standard deviations to rating-scale data in spite of the fact that these rating scales may fall short of the strict requirements for equal interval level of measurement; the arguments made by Harris and others above suggest that this common practice is not necessarily problematic.

This is another instance where actual researcher behaviors differ from the guidelines suggested in many statistics textbooks. If Stevens's level of measurement requirements are strictly enforced, statistics such as means and Pearson correlations could be computed only for variables that can be proved to have a true interval/ratio level of measurement. Unfortunately, many of the variables used in behavioral and social sciences (such as attitude-rating scales, personality trait measures, and so forth) probably fall somewhere between ordinal and interval in terms of level of measurement. A conservative interpretation of the guidelines about levels of measurement would lead to the conclusion that statistics such as mean and Pearson $r$ should not be applied to rating-scale data. However, in practice, many researchers do compute means, standard deviations, and Pearson correlations for variables that probably do not satisfy the strict requirements for equal interval level of measurement, such as attitude-rating scales. Thus, there is a discrepancy between the implicit standards for choice of statistics based on the theory of measurement that is still presented in many introductory statistics books and the data analysis practices of many researchers. When level of measurement falls into a gray area (where the information provided by scores is possibly better than ordinal but probably falls short of the requirements for a true interval level of measurement), many data analysts go ahead and apply statistics such as Pearson correlation and the $t$ test in spite of the fact that Stevens's measurement model is often interpreted as a prohibition of this practice. Stevens (1951) himself acknowledged that in some situations, violation of the measurement model can lead to reasonable results.

Tabachnick and Fidell (2007) and the other authors cited above have argued that it is more important to consider the distribution shapes for scores on quantitative variables (rather than their levels of measurement). Many of the statistical tests covered in introductory statistics books were developed based on assumptions that scores on quantitative variables are normally distributed. To evaluate whether a batch of scores in a sample has a nearly *normal distribution* shape, we need to know what an ideal normal distribution looks like. The next section reviews the characteristics of the standard normal distribution.

1.7 ♦ The Normal Distribution

Introductory statistics books typically present both empirical and theoretical distributions. An empirical distribution is based on scores from a sample, while a theoretical distribution is defined by a mathematical function or equation.

A description of an empirical distribution can be presented as a table of frequencies or in a graph such as a histogram. Sometimes it is helpful to group scores in order to obtain a more compact view of the distribution. SPSS® makes reasonable default decisions about grouping scores and the number of intervals and interval widths to use; these decisions