CHAPTER

19

Choosing the Right Statistics

Tools You Will Need

This chapter provides an organized overview for most of the statistical procedures presented in this book. The following items are considered background material for this chapter. If you doubt your knowledge of any of these items, you should review the appropriate chapter or section before proceeding.

- Descriptive statistics
- Mean (Chapter 3)
- Standard deviation (Chapter 4)
- Correlation (Chapter 15)
- Inferential Statistics (Chapters 9, 10, 11, 12, 13, 14, 15, 16, 17, 18)

Preview

After students have completed a statistics course, they occasionally are confronted with situations in which they have to apply the statistics they have learned. For example, in the context of a research methods course, or while working as a research assistant, students are presented with the results from a study and asked to do the appropriate statistical analysis. The problem is that many of these students have no idea where to begin. Although they have learned the individual statistics, they cannot match the statistical procedures to a specific set of data. Our goal for this chapter is to provide some help with the problem.

We assume that you know (or can anticipate) what your data look like. Therefore, we begin by presenting some basic categories of data so you can find that one that matches your own data. For each data category, we then present the potential statistical procedures and identify the factors that determine which are appropriate for you based on the specific characteristics of your data.

19.1 THREE BASIC DATA STRUCTURES

Most research data can be classified in one of three basic categories.

Category 1: A single group of participants with one score per participant.

Category 2: A single group of participants with two (or more) variables measured for each participant.

Category 3: Two (or more) groups of scores with each score a measurement of the same variable.

In this section, we present examples of each structure. Once you match your own data to one of the examples, you can proceed to the section of the chapter in which we describe the statistical procedures that apply to that example.

SCALES OF MEASUREMENT

Before we begin discussion of the three categories of data, there is one other factor that differentiates data within each category and helps to determine which statistics are appropriate. In Chapter 1, we introduced four scales of measurement and noted that different measurement scales allow different kinds of mathematical manipulation, which result in different statistics. For most statistical applications, however, ratio and interval scales are equivalent so we group them together for the following review.

Ratio scales and **interval scales** produce numerical scores that are compatible with the full range of mathematical manipulation. Examples include measurements of height in inches, weight in pounds, the number of errors on a task, and IQ scores.

Ordinal scales consist of ranks or ordered categories. Examples include classifying cups of coffee as small, medium, and large or ranking job applicants as 1st, 2nd, and 3rd.

Nominal scales consist of named categories. Examples include gender (male/female), academic major, or occupation.

Within each category of data, we present examples representing these three measurement scales and discuss the statistics that apply to each.

CATEGORY 1: A SINGLE GROUP OF PARTICIPANTS WITH ONE SCORE PER PARTICIPANT

This type of data often exists in research studies that are conducted simply to describe individual variables as they exist naturally. For example, a recent news report stated that half of American teenagers, ages 12 through 17, send 50 or more text messages a day. To get this number, the researchers had to measure the number of text messages for each individual in a large sample of teenagers. The resulting data consist of one score per participant for a single group.

It is also possible that the data are a portion of the results from a larger study examining several variables. For example, a college administrator may conduct a survey to obtain information describing the eating, sleeping, and study habits of the college's students. Although several variables are being measured, the intent is to look at them one at a time. For example, the administrator will look at the number of hours each week that each student spends studying. These data consist of one score for each individual in a single group. The administrator will then shift attention to the number of hours per day that each student spends sleeping. Again, the data consist of one score for each person in a single group. The identifying feature for this type of research (and this type of data) is that there is no attempt to examine relationships between different variables. Instead, the goal is to describe individual variables, one at a time.

Table 19.1 presents three examples of data in this category. Note that the three data sets differ in terms of the scale of measurement used to obtain the scores. The first set (a) shows numerical scores measured on an interval or ratio scale. The second set (b) consists of ordinal, or rank ordered categories, and the third set shows nominal measurements. The statistics used for data in this category are discussed in Section 19.2.

CATEGORY 2: A SINGLE GROUP OF PARTICIPANTS WITH TWO (OR MORE) VARIABLES MEASURED FOR EACH PARTICIPANT

These research studies are specifically intended to examine relationships between variables. Note that different variables are being measured, so each participant has two or more scores, each representing a different variable. Typically, there is no attempt to manipulate or control the variables; they are simply observed and recorded as they exist naturally.

Although several variables may be measured, researchers usually select pairs of variables to evaluate specific relationships. Therefore, we present examples showing pairs of variables. Table 19.2 presents four examples of data in this category. Once again, the four data sets differ in terms of the scales of measurement that are used. The first set of data (a) shows numerical scores for each set of measurements. For the second set (b), we have ranked the scores from the first set and show the resulting ranks. The third data set (c) shows numerical scores for one variable and nominal scores for the second variable. In the fourth set (d), both scores are measured on a nominal scale

TABLE 19.1

Three examples of data with one score per participant for one group of participants.

(a) Number of Text Messages Sent in Past 24 Hours	(b) Rank in Class for High School Graduation	(c) Got a Flu Shot Last Season
X	*X*	*X*
6	23rd	No
13	18th	No
28	5th	Yes
11	38th	No
9	17th	Yes
31	42nd	No
18	32nd	No

TABLE 19.2

Examples of data with two scores for each participant for one group of participants.

(a) SAT Score (*X*) and College Freshman GPA (*Y*)		(b) Ranks for the Scores in Set (a)	
X	*Y*	*X*	*Y*
620	3.90	7	8
540	3.12	3	2
590	3.45	6	5
480	2.75	1	1
510	3.20	2	3
660	3.85	8	7
570	3.50	5	6
560	3.24	4	4

(c) Age (*X*) and Wrist Watch Preference (*Y*)		(d) Gender (*X*) and Academic Major (*Y*)	
X	*Y*	*X*	*Y*
27	digital	M	Sciences
43	analog	M	Humanities
19	digital	F	Arts
34	digital	M	Professions
37	digital	F	Professions
49	analog	F	Humanities
22	digital	F	Arts
65	analog	M	Sciences
46	digital	F	Humanities

of measurement. The appropriate statistical analyses for these data are discussed in Section 19.3.

CATEGORY 3: TWO OR MORE GROUPS OF SCORES WITH EACH SCORE A MEASUREMENT OF THE SAME VARIABLE

A second method for examining relationships between variables is to use the categories of one variable to define different groups and then measure a second variable to obtain a set of scores within each group. The first variable, defining the groups, usually falls into one of the following general categories:

a. Participant characteristic: For example, gender or age.

b. Time: For example, before versus after treatment.

c. Treatment conditions: For example, with caffeine versus without caffeine.

If the scores in one group are consistently different from the scores in another group, then the data indicate a relationship between variables. For example, if the performance scores for a group of females are consistently higher than the scores for a group of males, then there is a relationship between performance and gender.

Another factor that differentiates data sets in this category is the distinction between independent-measures and repeated-measures designs. Independent-measures designs were introduced in Chapters 10 and 12, and repeated-measures designs were presented in Chapters 11 and 13. You should recall that an *independent-measures design,* also known as a *between-subjects design,* requires a separate group of participants for each group of scores. For example, a study comparing scores for males with scores for females would require two groups of participants. On the other hand, a *repeated-measures design,* also known as a *within-subjects design,* obtains several groups of scores from the same group of participants. A common example of a repeated-measures design is a before/after study in which one group of individuals is measured before a treatment and then measured again after the treatment.

Examples of data sets in this category are presented in Table 19.3. The table includes a sampling of independent-measures and repeated-measures designs as well as examples representing measurements from several different scales of measurement. The appropriate statistical analyses for data in this category are discussed in Section 19.4.

19.2 STATISTICAL PROCEDURES FOR DATA FROM A SINGLE GROUP OF PARTICIPANTS WITH ONE SCORE PER PARTICIPANT

One feature of this data category is that the researcher typically does not want to examine a relationship between variables but rather simply intends to describe individual variables as they exist naturally. Therefore, the most commonly used statistical procedures for these data are descriptive statistics that are used to summarize and describe the group of scores.

We should note that the same descriptive statistics used to describe a single group are also used to describe groups of scores that are a part of more complex data sets. For example, a researcher may want to compare a set of scores for males with a set of scores

TABLE 19.3

Examples of data comparing two or more groups of scores with all scores measuring the same variable.

(a) Attractiveness Ratings for a Woman in a Photograph Shown on a Red or a White Background

White	Red
5	7
4	5
4	4
3	5
4	6
3	4
4	5

(b) Performance Scores Before and After 24 Hours of Sleep Deprivation

Participant	Before	After
A	9	7
B	7	6
C	7	5
D	8	8
E	5	4
F	9	8
G	8	5

(c) Success or Failure on a Task for Participants Working Alone or in a Group

Alone	Group
Fail	Succeed
Succeed	Succeed
Succeed	Succeed
Succeed	Succeed
Fail	Fail
Fail	Succeed
Succeed	Succeed
Fail	Succeed

(d) Amount of Time Spent on Facebook (Small, Medium, Large) for Students from Each High School Class

Freshman	Sophomore	Junior	Senior
med	small	med	large
small	large	large	med
small	med	large	med
med	med	large	large
small	med	med	large
large	large	med	large
med	large	small	med
small	med	large	large

for females (data from category 3). However, the statistics used to describe the group of males would be the same as the descriptive statistics that would be used if the males were the only group in the study.

SCORES FROM RATIO OR INTERVAL SCALES: NUMERICAL SCORES

When the data consist of numerical values from interval or ratio scales, there are several options for descriptive and inferential statistics. We consider the most likely statistics and mention some alternatives.

Descriptive Statistics The most often used descriptive statistics for numerical scores are the mean (Chapter 3) and the standard deviation (Chapter 4). If there are a few extreme scores or the distribution is strongly skewed, the median (Chapter 3) may be better than the mean as a measure of central tendency.

Inferential Statistics If there is a basis for a null hypothesis concerning the mean of the population from which the scores were obtained, a single-sample *t* test (Chapter 9) can be used to evaluate the hypothesis. Some potential sources for a null hypothesis are as follows:

1. If the scores are from a measurement scale with a well-defined neutral point, then the *t* test can be used to determine whether the sample mean is significantly different from (higher than or lower than) the neutral point. On a 7-point rating scale, for example, a score of $X = 4$ is often identified as neutral. The null hypothesis would state that the population mean is equal to (greater than or less than) $\mu = 4$.
2. If the mean is known for a comparison population, then the *t* test can be used to determine whether the sample mean is significantly different from (higher than or lower than) the known value. For example, it may be known that the average score on a standardized reading achievement test for children finishing first grade is $\mu = 20$. If the sample consists of test scores for first grade children who are all the only child in the household, the null hypothesis would state that the mean for children in this population is also equal to 20. The known mean could also be from an earlier time, for example 10 years ago. The hypothesis test would then determine whether a sample from today's population indicates a significant change in the mean during the past 10 years.

The single-sample *t* test evaluates the statistical significance of the results. A significant result means that the data are very unlikely ($p < \alpha$) to have been produced by random, chance factors. However, the test does not measure the size or strength of the effect. Therefore, a *t* test should be accompanied by a measure of effect size such as Cohen's *d* or the percentage of variance accounted for, r^2.

SCORES FROM ORDINAL SCALES: RANKS OR ORDERED CATEGORIES

Descriptive Statistics Occasionally, the original scores are measurements on an ordinal scale. It is also possible that the original numerical scores have been transformed into ranks or ordinal categories (for example, small, medium, and large). In either case, the median is appropriate for describing central tendency for ordinal measurements and proportions can be used to describe the distribution of individuals across categories. For example, a researcher might report that 60% of the students were in the high self-esteem category, 30% in the moderate self-esteem category, and only 10% in the low self-esteem category.

Inferential Statistics If there is a basis for a null hypothesis specifying the proportions in each ordinal category for the population from which the scores were obtained, then a chi-square test for goodness of fit (Chapter 17) can be used to evaluate the hypothesis. With only two categories, the binomial test (Chapter 18) also can be used. For example, it may be reasonable to hypothesize that the categories occur equally often (equal proportions) in the population and the test would determine whether the sample proportions are significantly different. If the original data were converted from numerical values into ordered categories using z-score values to define the category boundaries, then the null hypothesis could state that the population distribution is normal, using proportions obtained from the unit normal table. A chi-square test for goodness of fit would determine whether the shape of the sample distribution is significantly different from a normal distribution. For example, the null hypothesis would state that the distribution has the following proportions, which describe a normal distribution according to the unit normal table:

$z < -1.5$	$-1.5 < z < -0.5$	$-0.5 < z < 0.5$	$0.5 < z < 1.5$	$z > 1.50$
6.68%	24.17%	38.30%	24.17%	6.68%

SCORES FROM A NOMINAL SCALE

For these data, the scores simply indicate the nominal category for each individual. For example, individuals could be classified as male/female or grouped into different occupational categories.

Descriptive Statistics The only descriptive statistics available for these data are the mode (Chapter 3) for describing central tendency or using proportions (or percentages) to describe the distribution across categories.

Inferential Statistics If there is a basis for a null hypothesis specifying the proportions in each category for the population from which the scores were obtained, then a chi-square test for goodness of fit (Chapter 17) can be used to evaluate the hypothesis. With only two categories, the binomial test (Chapter 18) also can be used. For example, it may be reasonable to hypothesize that the categories occur equally often (equal proportions) in the population. If proportions are known for a comparison population or for a previous time, the null hypothesis could specify that the proportions are the same for the population from which the scores were obtained. For example, if it is known that 35% of the adults in the United States get a flu shot each season, then a researcher could select a sample of college students and count how many got a shot and how many did not [see the data in Table 19.1(c)]. The null hypothesis for a chi-square test or a binomial test would state that the distribution for college students is not different from the distribution for the general population. For the chi-square test

	Flu Shot	No Flu Shot
H_0:	35%	65%

For the binomial test, H_0: $p = p(\text{shot}) = 0.35$ and $q = p(\text{no shot}) = 0.65$

Figure 19.1 summarizes the statistical procedures used for data in category 1.

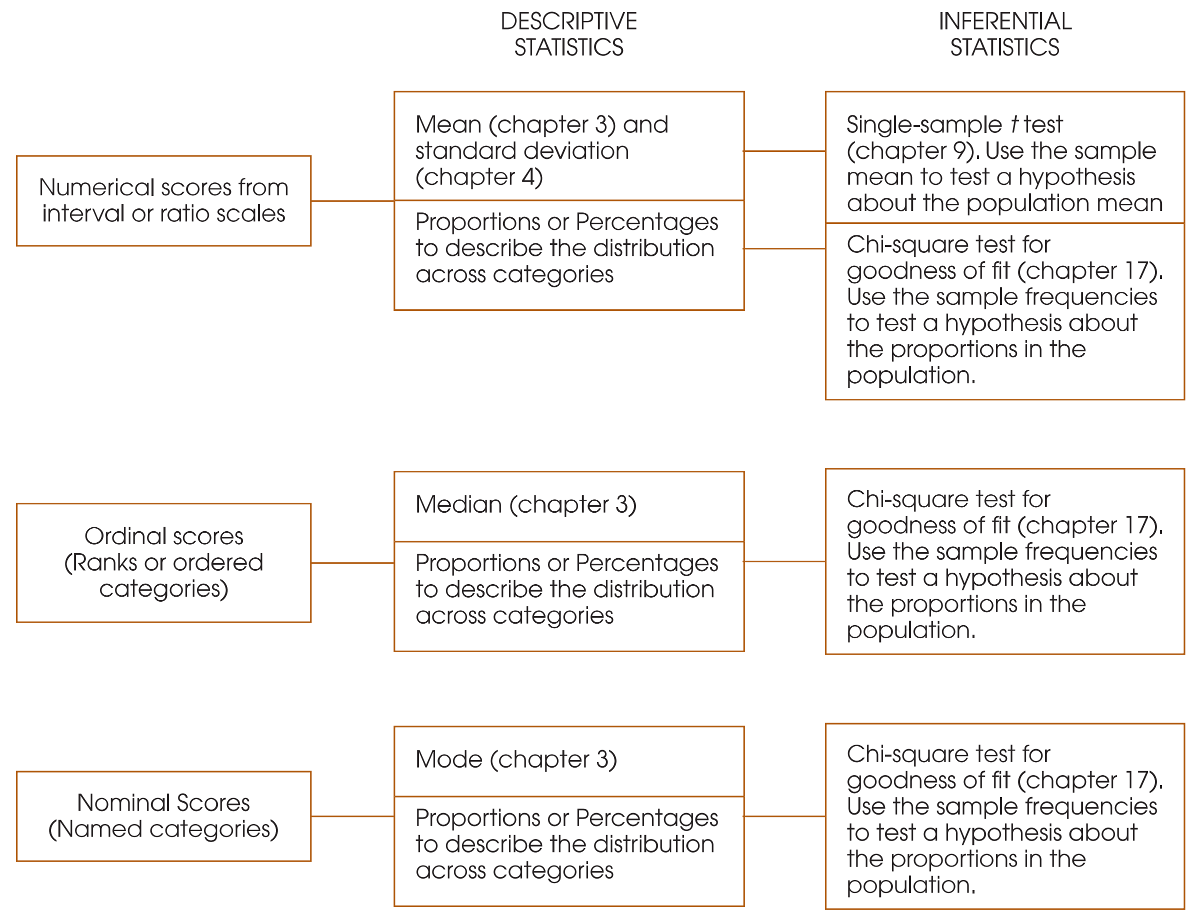

FIGURE 19.1

Statistics for Category 1 data. A single group of participants with one score per participant. The goal is to describe the variable as it exists naturally.

19.3 STATISTICAL PROCEDURES FOR DATA FROM A SINGLE GROUP OF PARTICIPANTS WITH TWO (OR MORE) VARIABLES MEASURED FOR EACH PARTICIPANT

The goal of the statistical analysis for data in this category is to describe and evaluate the relationships between variables, typically focusing on two variables at a time. With only two variables, the appropriate statistics are correlations (Chapter 15), linear regression (Chapter 16), and the chi-square test for independence (Chapter 17). With three or more variables, the applicable statistics are partial correlation (Chapter 15) and multiple regression (Chapter 16).

TWO NUMERICAL VARIABLES FROM INTERVAL OR RATIO SCALES

The Pearson correlation measures the degree and direction of linear relationship between the two variables (see Example 15.3 on p. 517). Linear regression determines the equation for the straight line that gives the best fit to the data points. For each X value in the data, the equation produces a predicted Y value on the line so that the squared distances between the actual Y values and the predicted Y values are minimized.

Descriptive Statistics The Pearson correlation serves as its own descriptive statistic. Specifically, the sign and magnitude of the correlation describe the linear relationship between the two variables. The squared correlation is often used to describe the strength of the relationship. The linear regression equation provides a mathematical description of the relationship between X values and Y. The slope constant describes the amount that Y changes each time the X value is increased by 1 point. The constant (Y intercept) value describes the value of Y when X is equal to zero.

Inferential Statistics The statistical significance of the Pearson correlation is evaluated by comparing the sample correlation with critical values listed in Table B6. A significant correlation means that it is very unlikely ($p < \alpha$) that the sample correlation would occur without a corresponding relationship in the population. Analysis of regression is a hypothesis-testing procedure that evaluates the significance of the regression equation. Statistical significance means that the equation predicts more of the variance in the Y scores than would be reasonable to expect if there were not a real underlying relationship between X and Y.

TWO ORDINAL VARIABLES (RANKS OR ORDERED CATEGORIES)

The Spearman correlation is used when both variables are measured on ordinal scales (ranks). If one or both variables consist of numerical scores from an interval or ratio scale, then the numerical values can be transformed to ranks and the Spearman correlation can be computed.

Descriptive Statistics The Spearman correlation describes the degree and direction of monotonic relationship; that is the degree to which the relationship is consistently one directional.

Inferential Statistics The statistical significance of the Spearman correlation is evaluated by comparing the sample correlation with critical values listed in Table B7. A significant correlation means that it is very unlikely ($p < \alpha$) that the sample correlation would occur without a corresponding relationship in the population.

ONE NUMERICAL VARIABLE AND ONE DICHOTOMOUS VARIABLE (A VARIABLE WITH EXACTLY 2 VALUES)

The point-biserial correlation measures the relationship between a numerical variable and a dichotomous variable. The two categories of the dichotomous variable are coded as numerical values, typically 0 and 1, to calculate the correlation.

Descriptive Statistics Because the point-biserial correlation uses arbitrary numerical codes, the direction of relationship is meaningless. However, the size of the correlation, or the squared correlation, describes the degree of relationship.

Inferential Statistics The data for a point-biserial correlation can be regrouped into a format suitable for an independent-measures t hypothesis test, or the t value can be computed directly from the point-biserial correlation (see the example on pages 542–544). The t value from the hypothesis test determines the significance of the relationship.

TWO DICHOTOMOUS VARIABLES

The phi-coefficient is used when both variables are dichotomous. For each variable, the two categories are numerically coded, typically as 0 and 1, to calculate the correlation.

Descriptive Statistics Because the phi-coefficient uses arbitrary numerical codes, the direction of relationship is meaningless. However, the size of the correlation, or the squared correlation, describes the degree of relationship.

Inferential Statistics The data from a phi-coefficient can be regrouped into a format suitable for a 2 × 2 chi-square test for independence, or the chi-square value can be computed directly from the phi-coefficient (see Example 17.3 on p. 613). The chi-square value determines the significance of the relationship.

TWO VARIABLES FROM ANY MEASUREMENT SCALES

The chi-square test for independence (Chapter 17) provides an alternative to correlations for evaluating the relationship between two variables. For the chi-square test, each of the two variables can be measured on any scale, provided that the number of categories is reasonably small. For numerical scores covering a wide range of value, the scores can be grouped into a smaller number of ordinal intervals. For example, IQ scores ranging from 93 to 137 could be grouped into three categories described as high, medium, and low IQ.

For the chi-square test, the two variables are used to create a matrix showing the frequency distribution for the data. The categories for one variable define the rows of the matrix and the categories of the second variable define the columns. Each cell of the matrix contains the frequency or number of individuals whose scores correspond to the row and column of the cell. For example, the gender and academic major scores in Table 19.2(d) could be reorganized in a matrix as follows:

	Arts	Humanities	Sciences	Professions
Female				
Male				

The value in each cell is the number of students with the gender and major identified by the cell's row and column. The null hypothesis for the chi-square test would state that there is no relationship between gender and academic major.

Descriptive Statistics The chi-square test is an inferential procedure that does not include the calculation of descriptive statistics. However, it is customary to describe the data by listing or showing the complete matrix of observed frequencies. Occasionally researchers describe the results by pointing out cells that have exceptionally large discrepancies. For example, in the Preview for Chapter 17 we described a study investigating eyewitness memory. Participants watched a video of an automobile accident and were questioned about what they saw. One group was asked to estimate the speed of the cars when they "smashed into" each other and another group was asked to estimate speed when the cars "hit" each other. A week later, they were asked additional questions, including whether they recalled seeing broken glass. Part of the description of the results focuses on cells reporting "Yes" responses. Specifically, the "smashed into" group had more than twice as many "Yes" responses than the "hit" group.

Inferential Statistics The chi-square test evaluates the significance of the relationship between the two variables. A significant result means that the distribution of frequencies in the data is very unlikely to occur ($p < \alpha$) if there is no underlying relationship between variables in the population. As with most hypothesis tests, a significant result does not provide information about the size or strength of the relationship. Therefore, either a phi-coefficient or Cramér's V is used to measure effect size.

THREE NUMERICAL VARIABLES FROM INTERVAL OR RATIO SCALES

To evaluate the relationship among three variables, the appropriate statistics are partial correlation (Chapter 15) and multiple regression (Chapter 16). A partial correlation measures the relationship between two variables while controlling the third variable. Multiple regression determines the linear equation that gives the best fit to the data points. For each pair of *X* values in the data, the equation produces a predicted *Y* value so that the squared distances between the actual *Y* values and the predicted *Y* values are minimized.

Descriptive Statistics A partial correlation describes the direction and degree of linear relationship between two variables while the influence of a third variable is controlled. This technique determines the degree to which the third variable is responsible for what appears to be a relationship between the first two. The multiple regression equation provides a mathematical description of the relationship between the two *X* values and *Y*. Each of the two slope constants describes the amount that *Y* changes each time the corresponding *X* value is increased by 1 point. The constant value describes the value of *Y* when both *X* values are equal to zero.

Inferential Statistics The statistical significance of a partial correlation is evaluated by comparing the sample correlation with critical values listed in Table B6 using $df = n - 3$ instead of the $n - 2$ value that is used for a routine Pearson correlation. A significant correlation means that it is very unlikely ($p < \alpha$) that the sample correlation would occur without a corresponding relationship in the population. Analysis of regression evaluates the significance of the multiple regression equation. Statistical significance means that the equation predicts more of the variance in the *Y* scores than would be reasonable to expect if there were not a real underlying relationship between the two *X*s and *Y*.

THREE VARIABLES INCLUDING NUMERICAL VALUES AND DICHOTOMOUS VARIABLES

Partial correlation (Chapter 15) and multiple regression (Chapter 16) can also be used to evaluate the relationship among three variables, including one or more dichotomous variables. For each dichotomous variable, the two categories are numerically coded, typically as 0 and 1, before the partial correlation or multiple regression is done. The *descriptive statistics* and the *inferential statistics* for the two statistical procedures are identical to those for numerical scores except that direction of relationship (or sign of the slope constant) is meaningless for the dichotomous variables.

Figure 19.2 summarizes the statistical procedures used for data in category 2.

19.4 STATISTICAL PROCEDURES FOR DATA CONSISTING OF TWO (OR MORE) GROUPS OF SCORES WITH EACH SCORE A MEASUREMENT OF THE SAME VARIABLE

Data in this category includes single-factor and two-factor designs. In a single-factor study, the values of one variable are used to define different groups and a second variable (the dependent variable) is measured to obtain a set of scores in each group. For a two-factor design, two variables are used to construct a matrix with the values of one variable defining the rows and the values of the second variable defining the columns. A third variable (the dependent variable) is measured to obtain a set of scores in each cell of the matrix. To simplify discussion, we focus on single-factor designs now and address two-factor designs in a separate section at the end of this chapter.

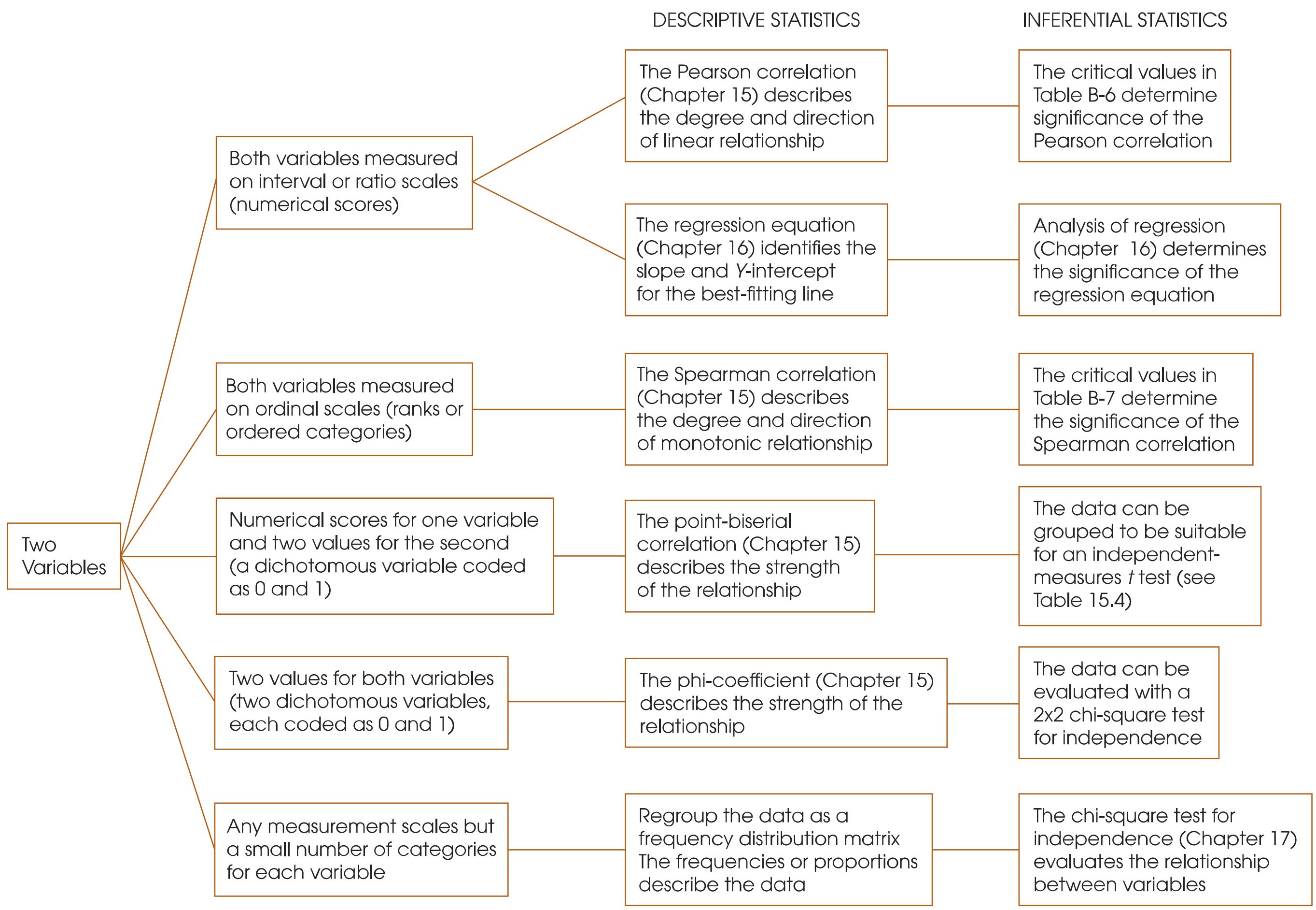

FIGURE 19.2

Statistics for Category 2 data. One group of participants with two (or more) variables measured for each participant. The goal is to describe and evaluate the relationship between variables.

The goal for a single-factor research design is to demonstrate a relationship between the two variables by showing consistent differences between groups. The scores in each group can be numerical values measured on interval or ratio scales, ordinal values (ranks), or simply categories on a nominal scale. The different measurement scales permit different types of mathematics and result in different statistical analyses.

SCORES FROM INTERVAL OR RATIO SCALES: NUMERICAL SCORES

Descriptive Statistics When the scores in each group are numerical values, the standard procedure is to compute the mean (Chapter 3) and the standard deviation (Chapter 4) as descriptive statistics to summarize and describe each group. For a repeated-measures study comparing exactly two groups, it also is common to compute the difference between the two scores for each participant and then report the mean and the standard deviation for the difference scores.

Inferential Statistics Analysis of variance (ANOVA) and *t* tests are used to evaluate the statistical significance of the mean differences between the groups of

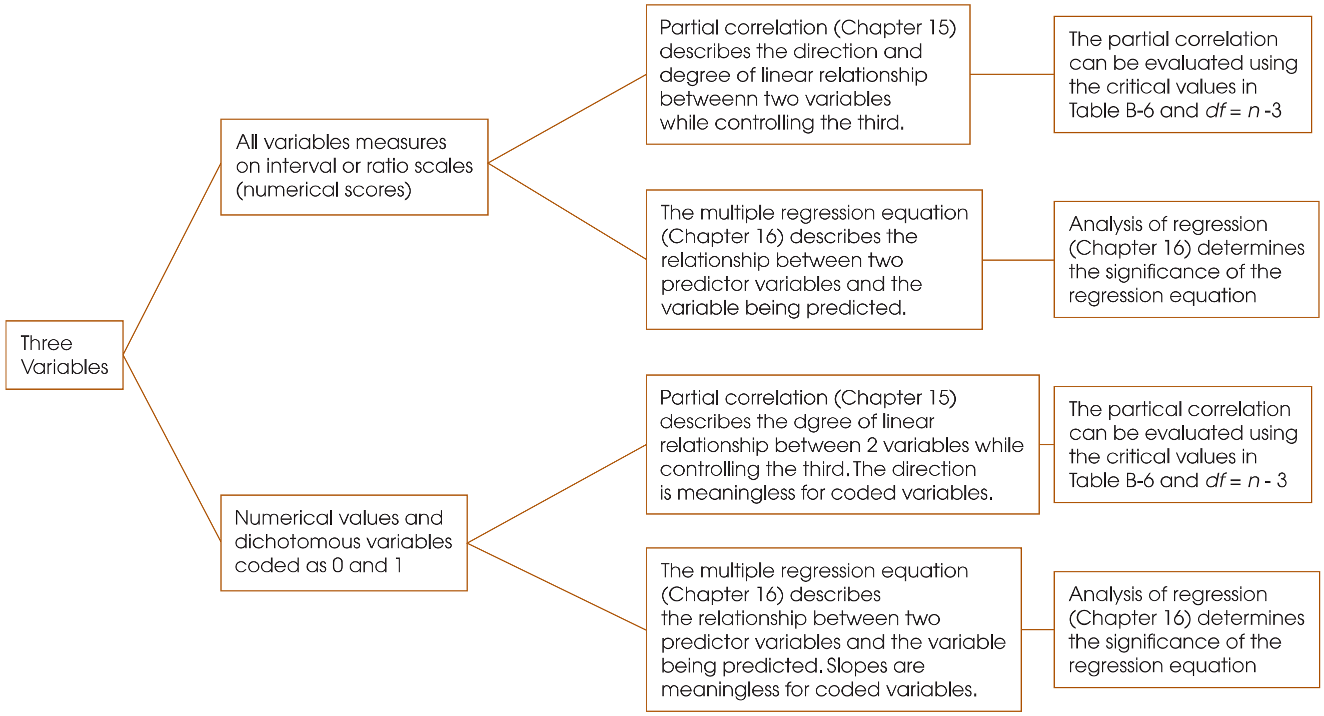

FIGURE 19.2—Cont'd

scores. With only two groups, the two tests are equivalent and either may be used. With more than two groups, mean differences are evaluated with an ANOVA. For independent-measures designs (between-subjects designs), the independent-measures t (Chapter 10) and independent-measures ANOVA (Chapter 12) are appropriate. For repeated-measures designs, the repeated-measures t (Chapter 11) and repeated-measures ANOVA (Chapter 13) are used. For all tests, a significant result indicates that the sample mean differences in the data are very unlikely ($p < \alpha$) to occur if there are not corresponding mean differences in the population. For an ANOVA comparing more than two means, a significant F-ratio indicates that post tests such as Scheffé or Tukey (Chapter 12) are necessary to determine exactly which sample means are significantly different. Significant results from a t test should be accompanied by a measure of effect size such as Cohen's d or r^2. For ANOVA, effect size is measured by computing the percentage of variance accounted for, η^2.

SCORES FROM ORDINAL SCALES: RANKS OR ORDERED CATEGORIES

For scores that are rank ordered, there are hypothesis tests developed specifically for ordinal data to determine whether there are significant differences in the ranks from one group to another. Also, if the scores are limited to a relatively small number of ordinal categories, then a chi-square test for independence can be used to determine whether there are significant differences in proportions from one group to another.

Descriptive Statistics Ordinal scores can be described by the set of ranks or ordinal categories within each group. For example, the ranks in one group may be consistently larger (or smaller) than ranks in another group. Or, the "large" ratings may be concentrated in one group and the "small" ratings in another.

Inferential Statistics Appendix E presents a set of hypothesis tests developed for evaluating differences between groups of ordinal data.

1. The Mann-Whitney *U* test evaluates the difference between two groups of scores from an independent-measures design. The scores are the ranks obtained by combining the two groups and rank ordering the entire set of participants from smallest to largest.
2. The Wilcoxon signed ranks test evaluates the difference between two groups of scores from a repeated-measures design. The scores are the ranks obtained by rank ordering the magnitude of the differences, independent of sign (+ or −).
3. The Kruskal-Wallis test evaluates differences between three or more groups from an independent-measures design. The scores are the ranks obtained by combining all of the groups and rank ordering the entire set of participants from smallest to largest.
4. The Friedman test evaluates differences among three or more groups from a repeated-measures design. The scores are the ranks obtained by rank ordering the scores for each participant. With three conditions, for example, each participant is measured three times and would receive ranks of 1, 2, and 3.

For all tests, a significant result indicates that the differences between groups are very unlikely ($p < \alpha$) to have occurred unless there are consistent differences in the population.

A chi-square test for independence (Chapter 17) can be used to evaluate differences between groups for an independent-measures design with a relatively small number of ordinal categories for the dependent variable. In this case, the data can be displayed as a frequency distribution matrix with the groups defining the rows and the ordinal categories defining the columns. For example, a researcher could group high school students by class (Freshman, Sophomore, Junior, Senior) and measure the amount of time each student spends on Facebook by classifying students into three ordinal categories (small, medium, large). An example of the resulting data is shown in Table 19.3(d). However, the same data could be regrouped into a frequency-distribution matrix as follows:

	Amount of Time Spent on Facebook		
	Small	Medium	Large
Freshman			
Sophomore			
Junior			
Senior			

The value in each cell is the number of students, with the high school class and amount of Facebook time identified by the cell's row and column. A chi-square test for independence would evaluate the differences between groups. A significant result indicates that the frequencies (proportions) in the sample data would be very unlikely ($p < \alpha$) to occur unless the proportions for the population distributions are different from one group to another.

SCORES FROM A NOMINAL SCALE

Descriptive Statistics As with ordinal data, data from nominal scales are usually described by the distribution of individuals across categories. For example, the scores in one group may be clustered in one category or set of categories and the scores in another group may be clustered in different categories.

Inferential Statistics With a relatively small number of nominal categories, the data can be displayed as a frequency-distribution matrix with the groups defining the rows and the nominal categories defining the columns. The number in each cell is the frequency, or number of individuals in the group, identified by the cell's row, with scores corresponding to the cell's column. For example, the data in Table 19.3(c) show success or failure on a task for participants who are working alone or working in a group. These data could be regrouped as follows:

	Success	Failure
Work Alone		
Work in a Group		

A chi-square test for independence (Chapter 17) can be used to evaluate differences between groups. A significant result indicates that the two sample distributions would be very unlikely ($p < \alpha$) to occur if the two population distributions have the same proportions (same shape).

TWO-FACTOR DESIGNS WITH SCORES FROM INTERVAL OR RATIO SCALES

Research designs with two independent (or quasi-independent) variables are known as two-factor designs. These designs can be presented as a matrix with the levels of one factor defining the rows and the levels of the second factor defining the columns. A third variable (the dependent variable) is measured to obtain a group of scores in each cell of the matrix (see Example 14.1 on page 477).

Descriptive Statistics When the scores in each group are numerical values, the standard procedure is to compute the mean (Chapter 3) and the standard deviation (Chapter 4) as descriptive statistics to summarize and describe each group.

Inferential Statistics A two-factor ANOVA is used to evaluate the significance of the mean differences between cells. The ANOVA separates the mean differences into three categories and conducts three separate hypothesis tests:

1. The main effect for factor *A* evaluates the overall mean differences for the first factor; that is, the mean differences between rows in the data matrix.
2. The main effect for factor *B* evaluates the overall mean differences for the second factor; that is, the mean differences between columns in the data matrix.
3. The interaction between factors evaluates the mean differences between cells that are not accounted for by the main effects.

For each test, a significant result indicates that the sample mean differences in the data are very unlikely ($p < \alpha$) to occur if there are not corresponding mean differences in the population. For each of the three tests, effect size is measured by computing the percentage of variance accounted for, η^2.

Figure 19.3 summarizes the statistical procedures used for data in category 3.

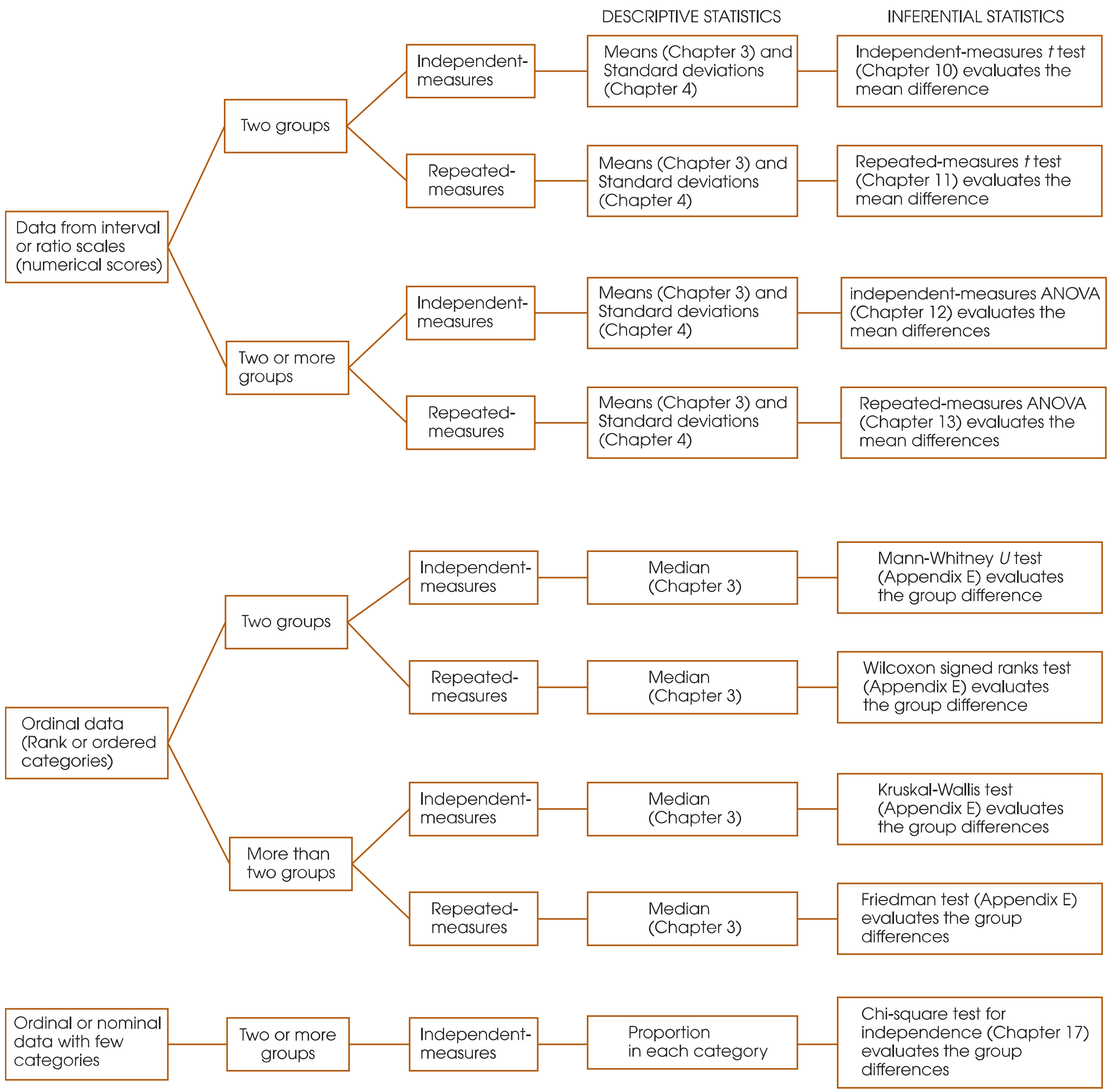

FIGURE 19.3

Statistics for Category 3 data. Two or more groups of scores with one score per participant. The goal is to describe and evaluate differences between groups of scores.

RESOURCES

Book Companion Website: www.cengage.com/psychology/gravettter

You can find a tutorial quiz and other learning exercises for Chapter 19 on the book companion website.

aplia

Improve your understanding of statistics with Aplia's auto-graded problem sets and immediate, detailed explanations for every question. To learn more, visit **www.aplia.com/statistics.**

CENGAGE brain .com

Log in to CengageBrain to access the resources your instructor requires. For this book, you can access:

Psychology CourseMate brings course concepts to life with interactive learning, study, and exam preparation tools that support the printed textbook. A textbook-specific website, Psychology CourseMate includes an integrated interactive eBook and other interactive learning tools including quizzes, flashcards, and more.

Visit www.cengagebrain.com to access your account and purchase materials.

PROBLEMS

Each problem describes a research situation and the data it produces. Identify the statistical procedures that are appropriate for the data. When possible, identify descriptive statistics, inferential statistics, and a measure of effect size.

1. Research suggests that the antioxidants in foods such as blueberries can reduce and even reverse age-related declines in cognitive functioning (Joseph et al., 1999). To test this phenomenon, a researcher selects a sample of $n = 25$ adults aged 70 to 75 and administers a cognitive function test to each participant. The participants then drink a blueberry supplement every day for 4 months before they are tested again. The researcher compares the scores before treatment with the scores after treatment to see if there is any change in cognitive function.
2. Recent budget cuts forced the city school district to increase class size in the elementary schools. To determine student reaction to the change, the district administered a survey to students asking whether the larger classes were "better, worse, or not different" from the classes the previous year. The results from the survey will be used to describe the students' attitude.
3. Last fall, the college introduced a peer-mentor program in which a sample of $n = 75$ freshmen was each assigned an upperclassman mentor. To evaluate the success of the program, the administration looked at the number of students who returned to the college to begin their second year. The data showed that 88% of the students in the peer-mentor program returned, compared to 72% for freshmen who were not in the program.
4. To examine the relationship between alcohol consumption and birth weight, a researcher selects a sample of $n = 20$ pregnant rats and mixes alcohol with their food for 2 weeks before the pups are born.

One newborn pup is randomly selected from each subject's litter and the average birth weight for the $n = 20$ pups is recorded. It is known that the average birth weight for regular rats (without exposure to alcohol) is $\mu = 5.6$ grams.

5. To examine the relationship between texting and driving skill, a researcher uses orange cones to set up a driving circuit in the high school parking lot. A group of students is then tested on the circuit, once while receiving and sending text messages and once without texting. For each student, the researcher records the number of orange cones hit while driving each circuit.

6. Childhood participation in sports, cultural groups, and youth groups appears to be related to improved self-esteem for adolescents (McGee, Williams, Howden-Chapman, Martin, & Kawachi, 2006). In a representative study, a researcher compares scores on a standardized self-esteem questionnaire for a sample of $n = 100$ adolescents with a history of group participation and a separate sample of $n = 100$ who have no history of group participation.

7. There is some evidence indicating that people with visible tattoos are viewed more negatively than people without visible tattoos (Resenhoeft, Villa, & Wiseman, 2008). In a similar study, a researcher showed male college students photographs of women and asked the students to rate the attractiveness of each woman using a 7-point scale. One of the women was selected as the target. For one group of participants, the target was photographed with a large tattoo on her shoulder and for a second group her photograph showed no tattoo. The researcher plans to compare the target's ratings for the two groups to determine whether the tattoo had any effect on perceived attractiveness.

8. A researcher investigated different combinations of temperature and humidity to examine how heat affects performance. The researcher compared three temperature conditions (70°, 80°, and 90°) with a high humidity and a low humidity condition for each temperature. A separate group of participants was tested in each of the six different conditions. For each participant, the researcher recorded the number of errors on a problem-solving task. The researcher would like to know how different combinations of temperature and humidity influence performance.

9. Hallam, Price, and Katsarou (2002) investigated the influence of background noise on classroom performance for children aged 10 to 12. In a similar study, students in one classroom worked on an arithmetic task with calming music in the background. Students in a second classroom heard aggressive, exciting music, and students in a third room had no music at all. The researchers measured the number of problems answered correctly for each student to determine whether the music conditions had any effect on performance.

10. A researcher is investigating the relationship between personality and birth order position. A sample of college students is classified into four birth-order categories (1st, 2nd, 3rd, 4th or later) and classified as being either extroverted or introverted.

11. A researcher is investigating the relationship between personality and birth order position. A sample of college students is classified into four birth-order categories (1st, 2nd, 3rd, 4th or later) and given a personality test that measures the degree of extroversion on a 50-point scale.

12. A survey of female high school seniors includes one question asking for the amount of time spent on clothes, hair, and makeup each morning before school. The researcher plans to use the results as part of a general description of today's high school students.

13. Brunt, Rhee, and Zhong (2008) surveyed 557 undergraduate college students to examine their weight status, health behaviors, and diet. In a similar study, researchers used body mass index (BMI) to classify a group of students into four categories: underweight, healthy weight, overweight, and obese. The students were also surveyed to determine the number of fatty and/or sugary snacks they eat each day. The researchers would like to use to data to determine whether there is a relationship between weight status and diet.

14. A researcher would like to determine whether infants, age 2 to 3 months, show any evidence of color preference. The babies are positioned in front of a screen on which a set of four colored patches is presented. The four colors are red, green, blue, and yellow. The researcher measures the amount of time each infant looks at each of the four colors during a 30 second test period. The color with the greatest time is identified as the preferred color for the child.

15. A researcher administers a survey to graduating seniors, asking them to rate their optimism about the current job market on a 7-point scale. The researcher plans to use the results as part of a description of today's graduating seniors.

16. Standardized measures seem to indicate that the average level of anxiety has increased gradually over the past 50 years (Twenge, 2000). In the 1950s, the average score on the Child Manifest Anxiety Scale was $\mu = 15.1$. A researcher administers the same test to a sample of $n = 50$ of today's children to determine whether there has been a significant change in the average anxiety level.

17. Belsky, Weinraub, Owen, and Kelly (2001) reported on the effects of preschool childcare on the development of young children. One result suggests that children who spend more time away from their mothers are more likely to show behavioral problems in kindergarten. Suppose that a kindergarten teacher is asked rank order the degree of disruptive behavior for the $n = 20$ children in the class.
  a. Researchers then separate the students into two groups: children with a history of preschool and children with little or no experience in preschool. The researchers plan to compare the ranks for the two groups.
  b. The researchers interview each child's parents to determine how much time the child spent in preschool. The children are then ranked according to the amount of preschool experience. The researchers plan to use the data to determine whether there is a relationship between preschool experience and disruptive behavior.

18. McGee and Shevlin (2009) found that an individual's sense of humor had a significant effect on how attractive the individual was perceived to be by others. In a similar study, female college students were given brief descriptions of three potential romantic partners. One was identified as the target and was described positively as being single, ambitious, and having good job prospects. For half of the participants, the description also said that he had a great sense of humor. For another half, it said that he has no sense of humor. After reading the three descriptions, the participants were asked to rank the three men 1st, 2nd, and 3rd in terms of attractiveness. For each of the two groups, the researchers recorded the number of times the target was placed in each ordinal position.

19. Numerous studies have found that males report higher self-esteem than females, especially for adolescents (Kling, Hyde, Showers, & Buswell, 1999). A recent study found that males scored an average of 8 points higher than females on a standardized questionnaire measuring self-esteem. The researcher would like to know whether this is a significant difference.

20. Research has demonstrated that IQ scores have been increasing, generation by generation, for years (Flynn, 1999). A researcher would like to determine whether this trend can be described by a linear equation showing the relationship between age and IQ scores. The same IQ test is given to a sample of 100 adults who range in age from 20 to 85 years. The age and IQ score are recorded for each person.

21. A researcher is investigating the effectiveness of acupuncture treatment for chronic back pain. A sample of $n = 20$ participants is obtained from a pain clinic. Each individual rates the current level of pain and then begins a 6-week program of acupuncture treatment. At the end of the program, the pain level is rated again and the researcher records whether the pain has increased or decreased for each participant.

22. Research results indicate that physically attractive people are also perceived as being more intelligent (Eagly, Ashmore, Makhijani, & Longo, 1991). As a demonstration of this phenomenon, a researcher obtained a set of $n = 25$ photographs of male college students. The photographs were shown to a sample of female college students who used a 7-point scale to rate several characteristics, including intelligence and attractiveness, for the person in each photo. The average attractiveness rating and the average intelligence rating were computed for each photograph. The researcher plans to use the averages to determine whether there is relationship between perceived attractiveness and perceived intelligence.

23. Research has shown that people are more likely to show dishonest and self-interested behaviors in darkness than in a well-lit environment (Zhong, Bohns, & Gino, 2010). In a related experiment, students were given a quiz and then asked to grade their own papers while the teacher read the correct answers. One group of students was tested in a well-lit room and another group was tested in a dimly-lit room. The researchers recorded the number of correct answers reported by each student to determine whether there was a significant difference between the two groups.

24. There is some evidence suggesting that you are likely to improve your test score if you rethink and change answers on a multiple-choice exam (Johnston, 1975). To examine this phenomenon, a teacher encouraged students to reconsider their answers before turning in exams. Students were asked to record their original answers and the changes that they made. When the exams were collected, the teacher found that 18 students improved their grades by changing answers and only 7 students had lower grades with the changes. The teacher would like to know if this is a statistically significant result.

25. A researcher is evaluating customer satisfaction with the service and coverage of three phone carriers. Each individual in a sample of $n = 25$ uses one carrier for 2 weeks, then switches to another for 2 weeks, and finally switches to the third for 2 weeks. Each participant then rates the three carriers.

a. Assume that each carrier was rated on a 10-point scale.

b. Assume that each participant ranked the three carriers 1st, 2nd and 3rd.

c. Assume the each participant simply identified the most preferred carrier of the three.

26. There is some research indicating that college students who use Facebook while studying tend to have lower grades than non-users (Kirschner & Karpinski, 2010). A representative study surveys students to determine the amount of Facebook use during the time they are studying or doing homework. Based on the amount of time spent on Facebook, students are classified into three groups (high, medium, and low time) and their grade point averages are recorded. The researcher would like to examine the relationship between grades and amount of time on Facebook.

27. To examine the effect of sleep deprivation on motor-skills performance, a sample of $n = 10$ participants was tested on a motor-skills task after 24 hours of sleep deprivation, tested again after 36 hours, and tested once more after 48 hours. The dependent variable is the number of errors made on the motor-skills task.

28. Ryan and Hemmes (2005) examined how homework assignments are related to learning. The participants were college students enrolled in a class with weekly homework assignments and quizzes. For some weeks, the homework was required and counted toward the student's grade. Other weeks, the homework was optional and did not count toward the student's grade. Predictably, most students completed the required homework assignments and did not do the optional assignments. For each student, the researchers recorded the average quiz grade for weeks with required homework and the average grade for weeks with optional homework to determine whether the grades were significantly higher when homework was required and actually done.

29. Ford and Torok (2008) found that motivational signs were effective in increasing physical activity on a college campus. In a similar study, researchers first counted the number of students and faculty who used the stairs and the number who used the elevators in a college building during a 30-minute observation period. The following week, signs such as "Step up to a healthier lifestyle" and "An average person burns 10 calories a minute walking up the stairs" were posted by the elevators and stairs and the researchers once again counted people to determine whether the signs had a significant effect on behavior.