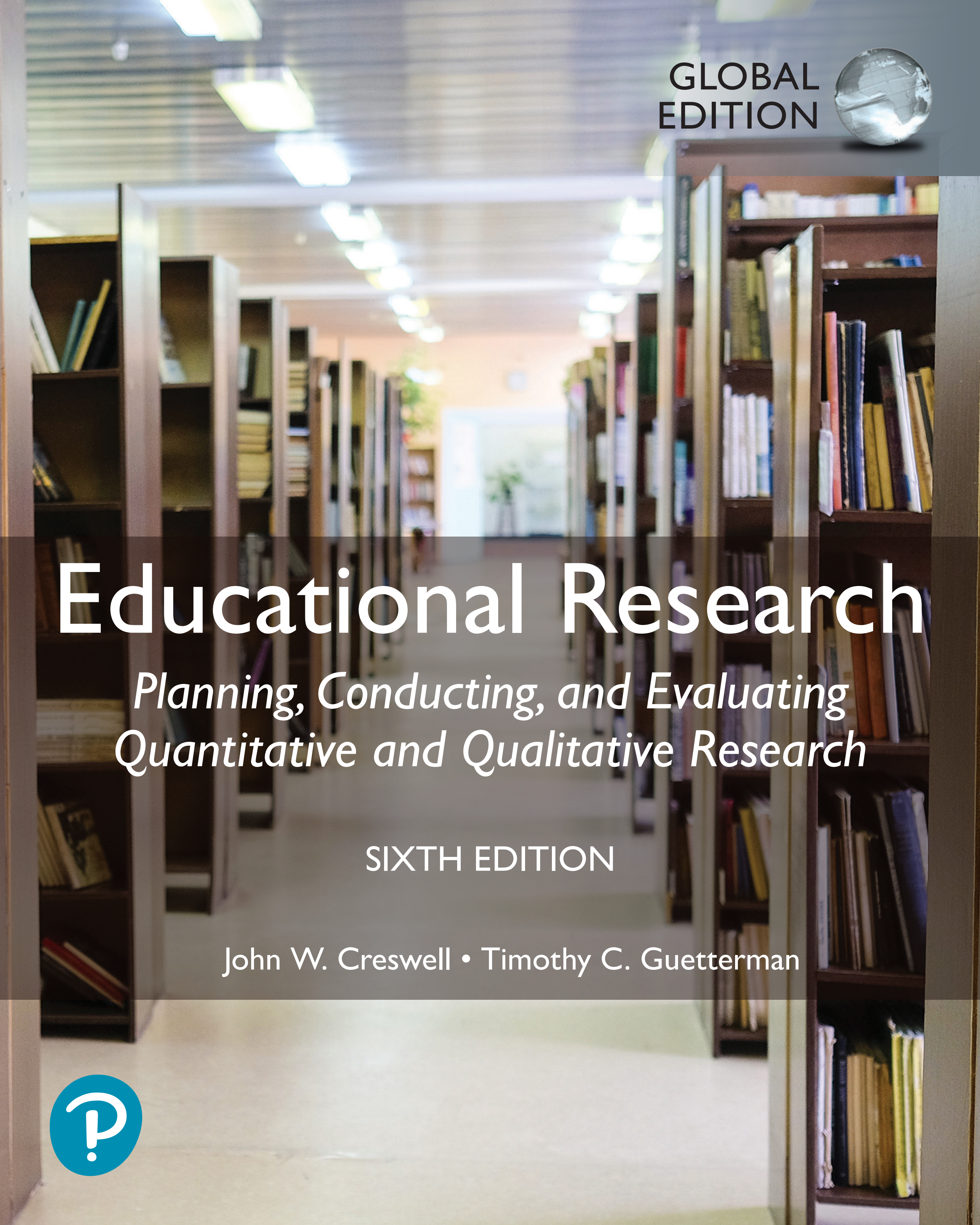


GLOBAL EDITION

Educational Research

Planning, Conducting, and Evaluating Quantitative and Qualitative Research

SIXTH EDITION

John W. Creswell • Timothy C. Guetterman

EDUCATIONAL
RESEARCH

data may result when instrument data are lost, individuals skip questions, participants are absent when you collect observational data, or individuals refuse to complete a sensitive question. For ethical reasons, you report how you handled missing data so that readers can accurately interpret the results (George & Mallery, 2016). Because these problems may occur, you need to clean the data and decide how to treat missing data.

Cleaning the Database

Cleaning the data is the process of inspecting the data for scores (or values) that are outside the accepted range. One way to accomplish this is by visually inspecting the data grid. For large databases, a frequency distribution (discussed shortly) will provide the range of scores to detect responses outside of acceptable ranges. For example, participants may provide a "6" for a *strongly agree* to *strongly disagree* scale when there are only five response options. Alternatively, a researcher might inadvertently enter an age of "73" in a study of elementary school age children when the only legitimate values are ages 5 to 12.

Another procedure is to use SPSS and have the program "sort cases" in ascending order for each variable. This process arranges the values of a variable from the smallest number to the largest, enabling you to easily spot out-of-range or misnumbered cases. Whatever the procedure, a visual inspection of data helps to clean the data and free them from visible errors before you begin the data analysis.

Assessing the Database for Missing Data

You need to examine your database for missing data. Missing data will yield fewer individuals to be included in the data analysis, and because you want as many people included in the analysis as possible, you need to correct as much as possible for missing data. **Missing data** are data that are missing in the database because participants do not supply them.

How should you handle missing data? The most obvious approach is to have a good instrument that individuals want to complete and are capable of answering so that missing data will not occur. In some research situations, you can contact individuals to determine why they did not respond. When individuals do not respond, something is wrong with your data collection, which may indicate faulty planning in your design.

You can expect, however, that questions will be omitted or that some participants will not supply information for whatever reason. In this case, you have a couple of options:

- You can eliminate participants with missing scores from the data analysis and include only those participants for whom complete data exist. This practice, in effect, may severely reduce the number of overall participants for data analysis.
- You can substitute numbers for missing data in the database for individuals. When the variable is categorical, this means substituting a value, such as "–9," that is easy to identify for all missing values in the data grid. When the variable is continuous (i.e., based on an interval scale), the process is more complex. Using SPSS, the researcher can have the computer program substitute a value for each missing score, such as an average number for the question for all study participants. You can substitute up to 15% of the missing data with scores without altering the overall statistical findings (George & Mallery, 2016). More advanced statistical procedures are also available for identifying substitute numbers for missing data (see Gall, Gall, & Borg, 2007).

MyLab Education Self-Check 6.2

MyLab Education Application Exercise 6.1: Preparing Data for Analysis

HOW DO YOU ANALYZE THE DATA?

After you prepare and organize the data, you are ready to analyze it. You analyze the data to address each one of your research questions or hypotheses. Questions or hypotheses in quantitative research require that you do the following:

- Describe trends in the data for a single variable or question on your instrument (e.g., "What is the self-esteem of middle school students?"). To answer this question, we need **descriptive statistics** that indicate general tendencies in the data (mean, median, and mode); the spread of scores (variance, standard deviation, and range); or a comparison of how one score relates to all others (z scores or percentile rank). We might seek to describe any of our variables: independent, dependent, control, or mediating.
- Compare two or more groups on the independent variable in terms of the dependent variable (e.g., "How do boys and girls compare in their self-esteem?"). To answer this question, we need **inferential statistics** in which we analyze data from a sample to draw conclusions about an unknown population. We assess whether the differences of groups (their means) or the relationship among variables is much greater or less than what we would expect for the total population if we could study the entire population.
- Relate two or more variables (e.g., "Does self-esteem relate to an optimistic attitude?"). To answer this question, we also use inferential statistics. Alternatively, we could test a hypothesis about the relationship of variables (e.g., "Self-esteem predicts an optimistic attitude among middle school children") using inferential statistics.
- Examine the relationship among variables through more advanced statistical procedures, such as regression analysis, meta-analysis, factors analysis, discriminant function analysis, path analysis, structural equation modeling or hierarchical linear modeling. These advanced statistical procedures are beyond the scope of this book but are introduced in Chapter 11.

Thus, we describe results to a single variable or question, or infer results from a sample to a population. In all quantitative research questions or hypotheses, we study individuals sampled from a population. However, in descriptive questions, we study only a single variable at a time; in inferential analysis, we analyze multiple variables at the same time. In addition, from comparing groups or relating variables, we can make predictions about the variables. We can test hypotheses that make predictions comparing groups or relating variables.

Conduct Descriptive Analysis

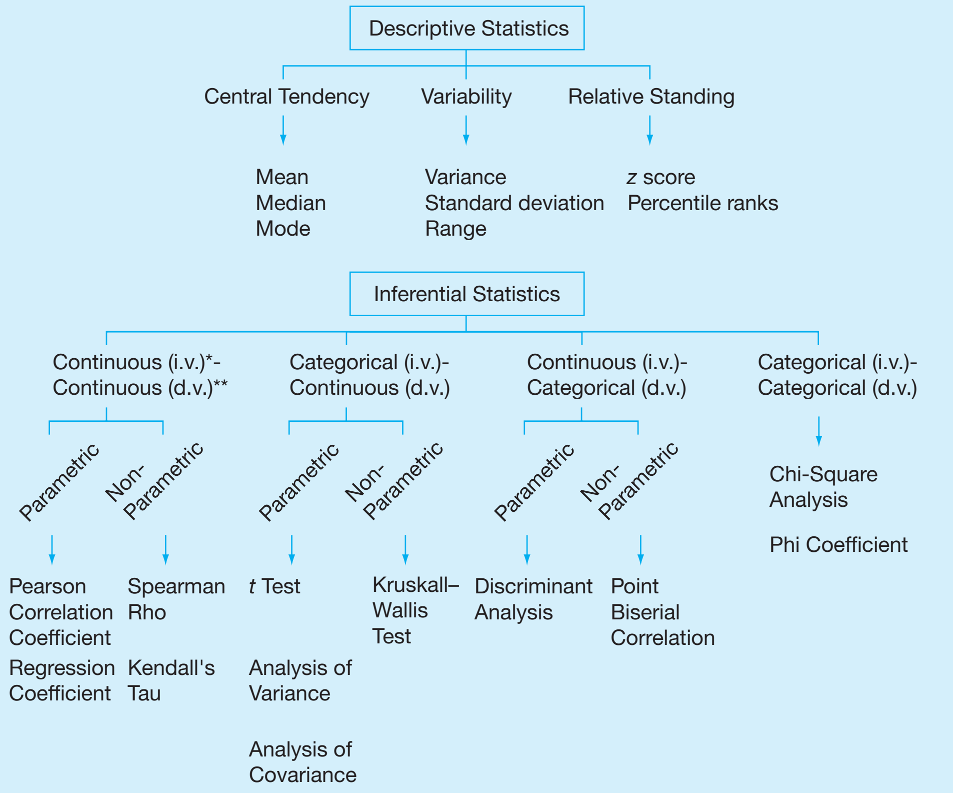
How do we analyze the data to describe trends? We use **statistics**, the calculations of values based on numbers. Many helpful books provide details about different statistics, their computation, and their assumptions (e.g., Field, 2013; Gravetter & Wallnau, 2013; Triola, 2018). We focus here on the statistics typically used in educational research.

Choosing a Descriptive Statistics Test

Descriptive statistics will help you summarize the overall trends or tendencies in your data, provide an understanding of how varied your scores might be, and provide insight into where one score stands in comparison with others. These three ideas are the central tendency, variability, and relative standing. Figure 6.3 portrays the statistical procedures that you can use to provide this information.

FIGURE 6.3

Examples of the Family of Statistics in Educational Research



* (i.v.) = independent variable
**(d.v.) = dependent variable

Measures of Central Tendency **Measures of central tendency** are summary numbers that represent a single value in a distribution of scores (Vogt & Johnson, 2016). They are expressed as an average score (the mean), the middle of a set of scores (the median), or the most frequently occurring score (the mode). In quantitative studies, researchers typically report all three measures. Table 6.3 portrays the differences between the three measures of central tendency for 10 students for whom we have depression scores.

The mean is the most popular statistic used to describe responses of all participants to items on an instrument. A **mean** (*M*) is the total of the scores divided by the number of scores. To calculate the mean, you sum all the scores and then divide the sum by the number of scores. In Table 6.3, you would divide the sum of 818 by 10 to get a mean of 81.80. In calculating other types of scores for other advanced statistics, the mean plays an important role. Notice that the scores in Table 6.3 are continuous and report a sample of 10 scores for depression. The mean gives us an average for all the scores.

We may want to know the middle score among all scores. This score is the median. The **median** score divides the scores, rank ordered from top to bottom, in half. Fifty percent of the scores lie above the median, and 50% lie below the median. To calculate

TABLE 6.3

Descriptive Statistics for Depression Scores

Raw Scores	*z* Score	Rank
60	−1.57	10.0
64	−1.28	20.0
75	−.49	30.0
76	−.42	50.0
76	−.42	50.0
83	+.09	60.0
93	+.81	70.0
94	+.92	80.0
98	+1.22	90.0
99	+1.24	100.0

Sum = 818

Mode = 76

Median = 79.5

Mean (M) = 81.8

Variance (standard deviation [SD^2]) = $\sum \frac{(\text{raw score} \; - M)^2}{N - 1}$

Variance = 193.29

$SD = \sqrt{\text{variance}}$

SD = 13.90

z score = raw score − M/SD

Range: minimum = 60; maximum = 99

this score, the researcher arrays all scores in rank order and then determines what score, the median, is halfway between all the scores. The median in Table 6.3 is halfway between 76 and 83, giving 79.5. There are five scores above 79.5 and five scores below it. Researchers often report the median score, but the score's usefulness is limited in this example. The median is more useful in situations with extreme outliers, such as income levels.

However, the mode provides useful information. The **mode** is the score that appears most frequently in a list of scores. It is used when researchers want to know the most common score in an array of scores on a variable. In Table 6.3, the most frequently reported score was 76, and it was a score for two people in the sample of 10. Researchers use the mode for reporting variables with categorical variables. Examine Table 6.4. Here is a categorical variable about the parental status. From looking at this table, we can see that married parents were more numerous than any other group (N = 22). This frequency distribution is also evident in Figure 6.4. The mode would be the "married" because they are represented more than any other category. Reporting the mean would report meaningless information. If we assigned numbers to each group (separated = 3, divorced = 2, and married = 1) and calculated a mean score, 86/50 = 1.72, it would

TABLE 6.4

Descriptive Statistics for the Categorical Variable "Parent Status"

	Parent Status			
	Frequency	Percent	Valid Percent	Cumulative Percent
Married	22	44.0	44.0	44.0
Divorced	20	40.0	40.0	84.0
Separated	13	16.0	16.0	100.00
Total	50	100.0	100.0	

not mean anything because no group is assigned this number. Thus, when we have categorical information, the mode reports meaningful information, but the mean does not. It is especially important to know what is meaningful because statistical programs will calculate the central tendency measures for you and you need decide which are useful.

Measures of Variability **Measures of variability** indicate the spread of the scores in a distribution. Range, variance, and standard deviation all indicate the amount of variability in a distribution of scores. This information helps us see how dispersed the responses are to items on an instrument. Variability also plays an important role in many advanced statistical calculations.

We can see how variable the scores are by looking at the range of scores. The **range of scores** is the difference between the highest and the lowest scores to items on an instrument. In Table 6.3, we see that the scores range from a low of 60 to a high of 99, a range of 39 points.

FIGURE 6.4

Bar Chart Depicting the Frequencies for the Categorical Variable "Parent Status"

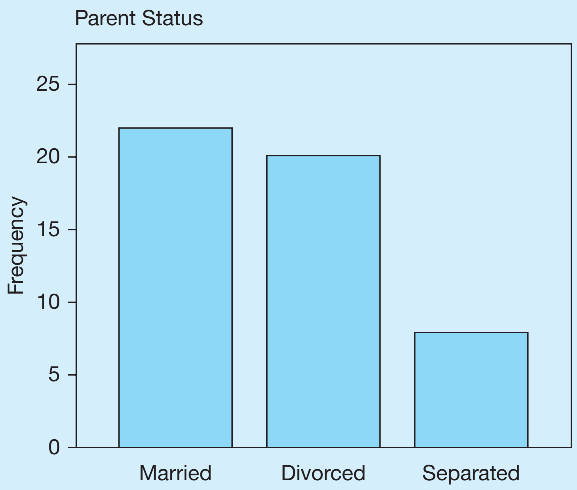
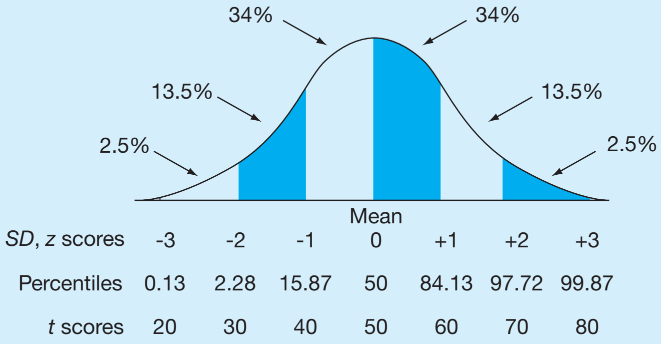


FIGURE 6.5

The Normal Curve



Source: Adapted from Gravetter and Wallnau (2013).

The **variance** indicates the dispersion of scores around the mean. Calculating this score is easy:

- Find the difference between the mean and the raw score for each individual
- Square this value for each individual
- Sum these squared scores for all individuals
- Divide by the total number of individuals minus 1

In our example, Table 6.3, the variance equals 193.29. This information, by itself, does not mean much, but it is a useful number when calculating statistics that are more advanced. The square root of the variance, the **standard deviation (*SD*)**, does provide useful information, and we look at it as an indicator of the dispersion or spread of the scores. In Table 6.3, the standard deviation is 13.90. If the scores had a standard deviation of 7.30, we would say that the variation around the mean is less than if the standard deviation is 13.90.

The meaning of the standard deviation becomes evident when we graph a theoretical distribution of scores, as shown in Figure 6.5. If we collected sample after sample of scores and plotted them on a graph, they would look like a bell-shaped curve (Figure 6.5). This is called a **normal distribution or normal probability curve**. In reality, the actual scores may not simulate the normal distribution (e.g., a distribution of salaries), but if we plotted the means of many samples, a normal curve would result. For example, if we generated 5,000 random samples and calculated a mean salary for each sample and then plotted these 5,000 means, the distribution would reflect a normal distribution. Looking again at Figure 6.5, the shaded areas indicate the percentage of scores likely to fall within each standard deviation from the mean. For example, 68% of the scores fall between +1 (34%) and −1 (34%) standard deviations from the mean; 95% between +2 (13.5 + 34) and −2 (13.5 + 34). You can also associate percentile scores and *z* scores with each standard deviation.

Percentiles provide another type of descriptive statistic. **Measures of relative standing** are statistics that describe one score relative to a group of scores. In Figure 6.5, 2.28% of the scores fall more than 2 standard deviations below the mean, and 97.72% of the scores are below the value 2 standard deviations above the mean. Knowing where a score falls in this distribution is a key factor in testing hypotheses.