Data Preparation and Analysis

CHAPTER

11

What You Should Know!

At this point, all the necessary steps have been established for finding a topic, developing a hypothesis, selecting a research design, and collecting data. The next step is to prepare the data for analysis and actually analyze the data. This analysis is where researchers will start to see the final results of their hard work. Chapters 11 and 12 focus on quantitative analysis, which is used more often than qualitative analysis in criminal justice and criminological studies. After completing this chapter, the reader should be able to:

1. Discuss the importance of data preparation.
2. Describe what occurs during the data coding process.
3. Describe the data-cleaning process.
4. Compare the different methods of dealing with missing data.
5. Explain why data may need to be recoded. Provide an example of how recoding may be done.
6. Identify and describe the three types of data analysis, including univariate, bivariate, and multivariate.
7. Recognize and explain the three types of statistical analysis, including descriptive, comparative, and inferential.
8. Present and discuss the four types of frequency distributions, including absolute frequency, relative frequency, cumulative frequency, and cumulative relative frequency.

9. Describe the various methods of visually presenting data in addition to frequency tables, including pie charts, bar charts, histograms, polygons, and line charts.
10. Discuss the three measures of central tendency, including mode, median, and mean.
11. Discuss the three measures of variability, including range, variance, and standard deviation.
12. Discuss what is meant by the terms *skewness* and *kurtosis*.

The first two sections of this textbook have provided a wealth of information regarding how to conduct research. Now it is time to address what to do with that data once they are collected. Following data collection, the next step is to prepare the data, which must occur before the data can be analyzed.

This chapter is not intended to teach students how to perform statistical analyses or serve as a replacement for a statistics class. The purpose of this chapter is to provide a general overview of preparing and analyzing data. Keep in mind that whenever referring to data in research, it should be used as a plural (datum is the singular version of data).

Data Preparation

Taking the time to properly prepare data for analysis is a vital step in the research process. Nevertheless, there is a tendency for researchers, even experienced researchers, to want to jump straight to the analysis step once they have the data. Will the results support the hypothesis? Has something new been discovered? Are these findings consistent with previous research? These are examples of questions one hopes to answer through the data analysis. However, not preparing the data for analysis would be like cooking with dirty utensils: You would have a meal at the end, but would you really want to eat it? The first set of tasks in data preparation generally consists of data coding, data entry, and data cleaning. After these tasks have been completed, then issues related to missing data and recoding of data can be addressed.

Data Coding

Coding data refers to the process of assigning values to the data for quantitative statistical analyses. The first step in coding is to determine the variable's level of measurement. In Chapter 6, we discussed the four levels of measurement: nominal, ordinal, interval, and ratio. In the data analysis process, interval-level variables are usually grouped with and treated as ratio-level variables. Keep in mind the questions for determining a variable's level of measurement. Are scores or categories the possible outcomes of a variable? If the possible outcomes are scores, then the variable is ratio. If the possible outcomes are categories, then the next question is whether the categories have an order. If the categories have

an order, then the variable is ordinal. If the categories do not have an order, then the variable is nominal.

Not all data require coding. Since ratio-level variables are based on scores, they can be left as is (e.g., age or income scores). Nominal- and ordinal-level variables are based on categories and require coding. For example, marital status (married, widowed, divorced, separated, never married) is a nominal-level variable. During the coding process each category is assigned a number, such as 1 for married, 2 for widowed, 3 for divorced, 4 for separated, and 5 for never married. Since marital status is a nominal-level variable, the order does not matter. However, the order is important when coding ordinal-level variables. For an ordinal-level variable such as socioeconomic status (low, middle, high), the lowest number should be associated with the lowest category. An example of coding for socioeconomic status would be 1 for low, 2 for middle, and 3 for high.

It is usually easier for researchers if the coding scheme is developed in advance. *Precoding* refers to a coding scheme that has been incorporated into the questionnaire or observational checklist. Such a data collection device may enable the researcher to enter data directly from the instrument rather than having to code the data first. The field research checklist in Chapter 10 (Figure 10-1) provides an example of precoding. Specifically, this form of precoding is known as *edge coding*, in reference to the responses lined up in the margin, which simplifies the data entry process.

During the coding process, qualitative data may also be converted into quantitative data. For example, an open-ended question about prior arrests may later be converted into numerical groupings based on the number of arrests, type of offenses, seriousness of the charges, conviction versus acquittal, or other logical groupings created by the researcher. Remember the previous warning about false precision when doing such a conversion. Be certain that there is an explainable logic to the numerical assignments.

Ultimately, coding decisions are up to the individual researcher, and researchers will develop their own style over time. One piece of advice is to keep the coding process as simple and straightforward as possible to avoid confusion. To ensure clarity, researchers should create codebooks or coding keys to show details related to the variables, such as the question (if part of a survey), level of measurement, and other coding details. **Figure 11-1** is an excerpt from the codebook for Dantzker's (2010) study of police pre-employment psychological screening.

Data Entry

Once the data are coded, they can then be entered into a computer software program dataset for analyses. **Figure 11-2** is an excerpt of the coded data after they have been entered into a dataset. In this figure, each column represents a

single variable and each row a single respondent. For example, if marital status was entered into a dataset, then the numbers 1–5 would be displayed in the marital status column. If respondent on line 1 was divorced, then 3 would be displayed on line one in the marital status column.

At various points in the text, it has been suggested that a computer can assist in conducting different aspects of the research process. The same is true for the analysis portion. Several statistical software packages are available to researchers, such as Excel, QuattroPro, Statistical Package for the Social Sciences (SPSS), and Statistical Analysis System. Each has its own quirks and specialties that require individuals to choose what best suits their research needs and abilities.

Regardless of which statistical software package is chosen, each allows for the entry, analyses, and storage of data. The authors recommend SPSS, which is available for personal computers after years of only being accessible through mainframes. This program provides the ability to conduct almost every type of statistical technique required, facilitated through a "point and click" interface that is less intimidating to new users than other statistical software packages. Before choosing one of the packages, researchers should examine their research design to determine which package will work best for their needs. Once the choice is made, it is simply a matter of learning how to use it and being able to interpret the results. Most statistical packages not only have handbooks but also excellent tutorials.

The key to data entry is accuracy. Typically, an individual will read and type the information from the data collection device (e.g., survey) into a dataset. Another possible strategy is to have a reliable person read the data aloud as it is entered into the computer to avoid having to continuously switch one's viewing from the code sheet to the computer screen. This process helps prevent mistakes. However, researchers should always review what has been entered to make certain that it is accurate.

Data entry can be time consuming and tedious. For many researchers, it is their least favorite part of the entire research process. It is not unheard of for university professors to assign graduate students this task rather than completing it themselves. However, there are several strategies that can ease the data entry process. The use of the Internet in data collection for surveys has greatly simplified the data entry process since the data is typically entered directly from the survey into a dataset, which can then be uploaded to one of the statistical software packages. If a telephone survey was conducted, it may have been possible to enter the data directly into the computer as the questions were asked. Another shortcut is to use optica-scan sheets, such as the Scantron sheets used in taking multiple-choice examinations. This strategy permits the data to be entered directly from the sheets marked by the respondents. Regardless of which technique is used, the data need to be cleaned before analysis can begin.

Data Cleaning

Data cleaning is the process of examining and reviewing the data for errors (Adler & Clark, 2007; Frankfort-Nachmias & Leon-Guerrero, 2008; Shaughnessy, Zechmeister, & Zechmeister, 2008). As in data entry, data cleaning can be tedious but is a crucial important step since much of the data entry in research is still done by hand. Mistakes that might have occurred during the initial recording of data or data entry are corrected, resulting in data that have been "cleaned." The first step in data cleaning is reviewing the data that has been entered for accuracy.

If a researcher is using a computer program that is programmed to check for errors, it may either beep or refuse to accept data that does not meet the coding requirements for that variable. This process is referred to as *automatic data cleaning*. For example, assume that a Likert scale was used where 1 represents strongly agree, 2 agree, 3 neither agree nor disagree, 4 disagree, and 5 strongly disagree. If the researcher tries to enter a value of 6, the program may beep as an alert of the error.

Not all statistical software packages have this automatic function. Therefore, cleaning must be done by the researcher. In addition to visually checking the data, creating a frequency table for each variable can be helpful. In a frequency table, all scores and categories are displayed for the selected variable, which allows the researcher to quickly see if there are errors. Using the frequency table, any incorrectly entered value should be easily observable.

Another type of data cleaning is contingency cleaning. In this technique, review of the data is expedited by the knowledge that certain responses should only have been made by certain individuals. Not all females would indicate having given birth to a child, but no males should have indicated such. As one is reviewing the data, there are certain questions that logically lead to similar responses on following questions. If they do not match the expectation, then that particular questionnaire or form should be reviewed to check the accuracy of what was entered. After the data have been cleaned, the next step is to determine how to handle missing data.

Missing Data

It is not unusual for data to be "lost" during the data collection process. Missing data may be caused by an oversight in data entry, which can usually be easily corrected during the data cleaning process by obtaining the right information from the survey form or code sheet. It also may occur because the respondents accidentally overlooked or deliberately chose not to answer a particular question. One strategy to decrease the likelihood of missing data from occurring is to provide an "escape" response, such as "no response" or "does not apply." These responses will not provide information applicable to the actual variable but are preferable to having missing data. If missing data does occur, there are several accepted strategies for handling this issue.

If the data have been left out on a single question, researchers may choose to enter it as a nonresponse. To do so, a value that is not being used within the variable itself is selected to represent a nonresponse. For example, the marital status variable has valid categories numbered 1 through 5. Therefore, 0 could be used to indicate a nonresponse. However, researchers frequently use a value that differs from the valid responses so that the nonresponse is not confused with the valid responses. Instead of 0, researchers may have used 99 (assuming the researchers are not using continuous variables in which 99 could be a possible response) to indicate a nonresponse. This strategy is the preferred method of dealing with nonresponses.

Another option for dealing with nonresponses is to assume that the missing data are caused by an oversight rather than an intentional omission. In this situation, researchers might look at the other responses to try to determine what the missing response would most likely have been. For example, if on related questions a respondent had indicated support for strict law enforcement, one might assume a similar answer on the unanswered question and enter this information into the dataset. Additionally, if the instrument used is a Likert scale, then the nonresponse can be classified as a "neither agree nor disagree" or "do not know" (depending on how the scale is worded). Although some researchers use these strategies, it is not recommended because it can lead to challenges regarding the objectivity and validity of the analysis.

Yet another option is to exclude from the analysis the data collection instrument containing the omission. If respondent 1 did not answer one of the questions, then all of his or her responses are eliminated from the analysis. This option may be appropriate if respondent 1 did not answer multiple questions. However, if only one or two questions are not answered, then this solution can lead to the loss of worthwhile information.

Recoding Data

Depending on the goal of the analysis, it may be necessary to recode certain variables. For example, using a ratio-level age variable to create a frequency table or a cross-tabulation (**Table 11-1**) may result in a table too large to be useful due to the large number of possible scores. A comparison of age by income could result in a table that is lengthy and confusing. Assuming that age was presented in columns and income in rows, if the people surveyed ranged from 20 years of age up to 70 years of age, and if their incomes ranged from $10,000 to $100,000, the resulting table would be enormous. Because each individual year of age would have to be represented, there would be 51 columns in this hypothetical table. Because there would likely be as many different incomes as there were respondents, the table could end up with thousands of rows. By collapsing the categories into logical groupings, the data can be presented clearly and concisely (Table 11-1).

TABLE 11-1

Hypothetical Age by Income Comparison

(Ordinal Data) Age Income	20–29	30–39	40–49	50–59	60–69	70+
$10,000 to $19,000	50	30	20	10	05	15
$20,000 to $29,999	100	80	50	30	50	80
$30,000 to $39,999	80	100	80	60	80	60
$40,000 to $49,000	50	100	160	120	100	80
$50,000 to $59,000	30	80	120	150	120	50
$60,000 to $69,000	20	50	100	120	100	40
$70,000 to $79,999	10	40	80	100	80	30
$80,000 to $89,000	05	30	60	80	60	20
$90,000 to $99,000	05	20	40	60	40	10
$100,000+	00	10	30	50	30	10

The scores or categories within an existing variable can be grouped to create a new variable. A ratio-level variable can be recoded into a nominal- or ordinal-level variable. For example, scores in a ratio-level age variable (e.g., 28, 30, 45, 62) can be recoded into an ordinal-level variable with three categories: less than 30, 30–50, greater than 50. An ordinal-level variable can be recoded into a different ordinal-level variable or into a nominal-level variable. A nominal-level variable can only be recoded into a different nominal-level variable. For example, a nominal-level variable based on vehicle type may have four categories: cars, trucks, SUVs, and vans. This variable could be recoded into a different nominal-level variable based on whether the vehicle is a car with the possible response categories of yes or no. Scores and categories can be grouped to create a new variable, but existing categories cannot be subdivided to create a new variable.

Data Analysis

Now that the data have been entered into a dataset and cleaned, analysis of the data may begin. There are three types of data analyses: (1) univariate, (2) bivariate, and (3) multivariate. With univariate analysis, one variable at a time is examined (Frankfort-Nachmias & Leon-Guerrero, 2008; Gavin, 2008; Walker & Maddan, 2009). Descriptive statistics are used to analyze individual variables and may include frequencies and measures of central tendency and variability. Descriptive statistics are discussed in detail later in this chapter.

Bivariate analysis is when the relationship between two variables is examined. ...tive or inferential depending on the nature of

TABLE 11-1

Hypothetical Age by Income Comparison

(Ordinal Data) Age Income	20–29	30–39	40–49	50–59	60–69	70+
$10,000 to $19,000	50	30	20	10	05	15
$20,000 to $29,999	100	80	50	30	50	80
$30,000 to $39,999	80	100	80	60	80	60
$40,000 to $49,000	50	100	160	120	100	80
$50,000 to $59,000	30	80	120	150	120	50
$60,000 to $69,000	20	50	100	120	100	40
$70,000 to $79,999	10	40	80	100	80	30
$80,000 to $89,000	05	30	60	80	60	20
$90,000 to $99,000	05	20	40	60	40	10
$100,000+	00	10	30	50	30	10

The scores or categories within an existing variable can be grouped to create a new variable. A ratio-level variable can be recoded into a nominal- or ordinal-level variable. For example, scores in a ratio-level age variable (e.g., 28, 30, 45, 62) can be recoded into an ordinal-level variable with three categories: less than 30, 30–50, greater than 50. An ordinal-level variable can be recoded into a different ordinal-level variable or into a nominal-level variable. A nominal-level variable can only be recoded into a different nominal-level variable. For example, a nominal-level variable based on vehicle type may have four categories: cars, trucks, SUVs, and vans. This variable could be recoded into a different nominal-level variable based on whether the vehicle is a car with the possible response categories of yes or no. Scores and categories can be grouped to create a new variable, but existing categories cannot be subdivided to create a new variable.

Data Analysis

Now that the data have been entered into a dataset and cleaned, analysis of the data may begin. There are three types of data analyses: (1) univariate, (2) bivariate, and (3) multivariate. With univariate analysis, one variable at a time is examined (Frankfort-Nachmias & Leon-Guerrero, 2008; Gavin, 2008; Walker & Maddan, 2009). Descriptive statistics are used to analyze individual variables and may include frequencies and measures of central tendency and variability. Descriptive statistics are discussed in detail later in this chapter.

Bivariate analysis is when the relationship between two variables is examined. This examination may be comparative or inferential depending on the nature of

the study. If a researcher is exploring the relationship between the two variables, then comparative statistics are used. Comparative statistics usually involve an analysis of the characteristics of the two variables to better describe the relationship between them. Even though the relationship between two variables is being examined, the variables themselves are not being labeled as independent and dependent variables. In other words, comparative statistics does not try to indicate the influence one variable has on another variable. Depiction of crime rates over time, noting the percentage of change in criminal activity, and trend analyses are examples of comparative statistics. Comparative statistics exist in a gray area between descriptive and inferential statistics and are categorized as a separate type of statistics here, according to the authors' practice. If the hypothetical Table 11-1 was real, it would be an example of comparative statistics. If one of the variables is identified as being dependent and the other variable is identified as being independent, then the analysis of their relationship becomes inferential. Inferential statistics indicates that the researcher is trying to use the knowledge gained from the inferential statistics to make predictions or inferences about the relationship between the two variables. Specifically, inferential statistics are used to predict the outcome of the dependent variable based on the influence of the independent variable.

The final type of statistical analysis is multivariate analysis. Multivariate analysis is the examination of three or more variables. This technique is inferential in nature in that descriptive and comparative statistical analyses (through univariate and bivariate analyses) of the data have already been conducted and now the focus is on examining the relationships among several variables. From this examination, a researcher is trying to develop explanations for the observed relationships. Inferential statistics are covered in detail in Chapter 12.

Statistical Analysis

Statistics are data presented in a manner that best represents what it is the researcher wants to discuss. Statistical analyses might best be viewed as processes for problem solving (Frankfort-Nachmias & Leon-Guerrero, 2008; Gavin, 2008; Walker & Maddan, 2009). Statistics are used in criminal justice and criminology to help describe a variety of associated aspects, such as crime rates, number of police officers, or prison populations. In addition, statistics can be used to make inferences about a phenomenon. As discussed in the preceding section, there are three types of statistics: (1) descriptive, whose function is describing the data; (2) comparative, whose function is to compare the attributes (or characteristics) of two variables; and (3) inferential, whose function is to make an inference, estimation, or prediction about the data.

The remainder of this chapter focuses on providing an overview of the processes and terminology of descriptive statistics. Understanding the

basic principles is important so that one can comprehend the work of other researchers.

Descriptive Statistics

When researchers are interested in knowing selected characteristics about their sample, they require some form of descriptive statistics. One of the most common descriptive statistics is the use of frequencies. When data are first collected, they are referred to as *raw data*. In other words, they have not been prepared and are not neatly organized. A first step to organizing the data, after coding and entry, is through a frequency distribution, which is commonly used to describe sample characteristics. The descriptive data in Table 11-2 come from Dantzker's (2010) study. These data describe the sample's characteristics of interest.

Frequency Distributions

Frequency distributions simply indicate the number of times a particular score or characteristic occurs in the sample, which can be reported in whole numbers

TABLE 11-2

From the Real World: Sample Demographics

	Police psychologists	Clinical psychologists	Total
Gender			
Female	11 (31)[a] (50)[b]	11 (28) (50)	22 (29) (100)
Male	25 (69) (47)	28 (72) (53)	53 (71) (100)
Total	**36 (100) (48)**	**39 (100) (52)**	75 (100) (100)
Years providing police services			
<1 years	1 (2) (33)	2 (5) (67)	3 (4) (100)
1–5 years	6 (16) (40)	9 (23) (60)	15 (20) (100)
6–10 years	10 (27) (76)	3 (7) (24)	13 (17) (100)
>10 years	20 (55) (44)	25 (65) (56)	45 (60) (100)
Total	**37 (100) (47)**	39 (100) (53)	76 (100) (100)
Agency types served			
Municipal	5 (14) (29)	12 (32) (71)	17 (23) (100)
County	0	7 (18) (100)	7 (9) (100)
State	1 (3) (25)	3 (8) (75)	4 (5) (100)
Combo	31 (83) (66)	16 (42) (34)	47 (63) (100)
Total	**37 (100) (49)**	38 (100) (51)	75 (100) (100)

[a]Column Percentage [b]Row Percentage

...employment psychological screening.

and percentages. There are four types of frequency distributions: (1) absolute, (2) relative, (3) cumulative, and (4) cumulative relative (Frankfort-Nachmias & Leon-Guerrero, 2008; Gavin, 2008; Walker & Maddan, 2009). Each offers a statistically sound indication of the sample's composition. Absolute and relative frequencies are the most reported. It is recommended that the type that seems to be the most desirable for reporting the data contained in the research be selected.

Absolute frequency distributions (also known simply as *frequencies*) display the data based on the number of times a score or category appears. Relative frequency distributions (also referred to as *percentages*) are the percentage equivalent of absolute frequency distributions and are determined by dividing the absolute frequency by the total number of cases. When dealing with large numbers, it is easier for readers to interpret percentages than raw numbers. Cumulative frequency distributions enable readers to see what the products of the grouping are in the frequencies table by adding the absolute frequency of each previous category or score. Cumulative relative frequency distributions do the same as cumulative frequency distributions but add relative frequencies rather than numbers. Table 11-3 demonstrates how the four types of frequencies are related.

Displaying Frequencies

Frequencies are just one means of describing the data. Other means include pie charts, bar graphs, histograms and polygons, line charts, and maps. As

TABLE 11-3

Absolute, Relative, Cumulative, and Cumulative Relative Frequency Distributions of Targeted Sample of Psychologist Members of APA by Region

	Absolute frequency	Relative frequency	Cumulative frequency	Cumulative relative frequency
EN Central	531	14.7	531	14.7
ES Central	176	4.9	707	19.6
Mid Atlantic	672	18.6	1379	38.2
Mountain	210	5.8	1589	44
New England	351	9.7	1940	53.8
Pacific	607	16.8	2547	70.6
South Atlantic	650	18	3197	88.6
WN Central	221	6.1	3418	

demonstrated in **Figure 11-3** (a pie chart), the frequencies or percentages can be depicted in a simplified picture form. Such a display is clear and easily interpreted. Bar charts are also easily interpreted (**Figure 11-4**). Pie and bar charts are used to display nominal- and ordinal-level data. Histograms are visually similar to bar graphs but are used to display interval- and ratio-level data (**Figure 11-5**). To indicate the continuous nature of the variable, the bars are adjacent to each other. In a bar chart, there is space between each bar, which indicates a categorical rather than continuous variable. Polygons display the same data as histograms but use dots instead of bars. Lines are then drawn between the dots to reveal the shape of the distribution. **Figure 11-6** is an example of a frequency polygon.

In addition to these frequency-presentation techniques, criminal justice and criminological researchers may also use line charts. Line charts are polygons that demonstrate scores across time (**Figure 11-7**). As such, they are useful to reflect changes over time in the same dot-and-line format as frequency polygons.

In addition to the tables, graphs, and charts discussed previously, there are four other ways to describe the properties of the data: (1) measures of central tendency, (2) measures of variability, (3) skewness, and (4) kurtosis.

Measures of Central Tendency

Because frequencies can be quite cumbersome, researchers sometimes require a way to summarize the data in a simpler manner. Measures of central tendency are

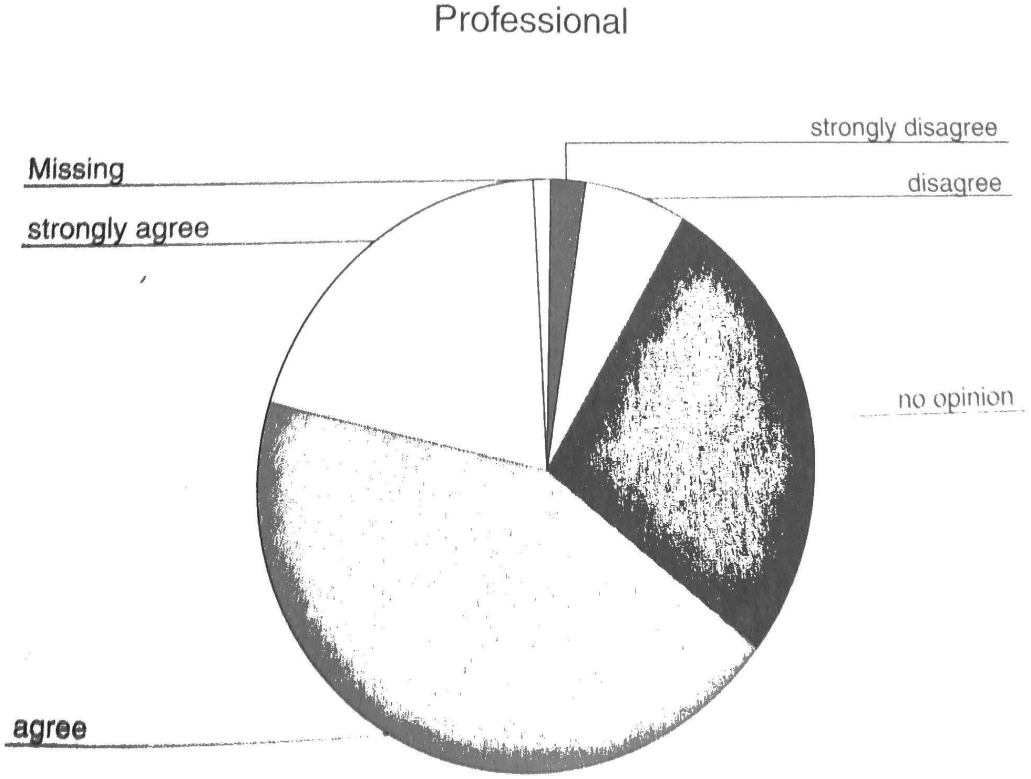

Figure 11-3 Pie Chart.

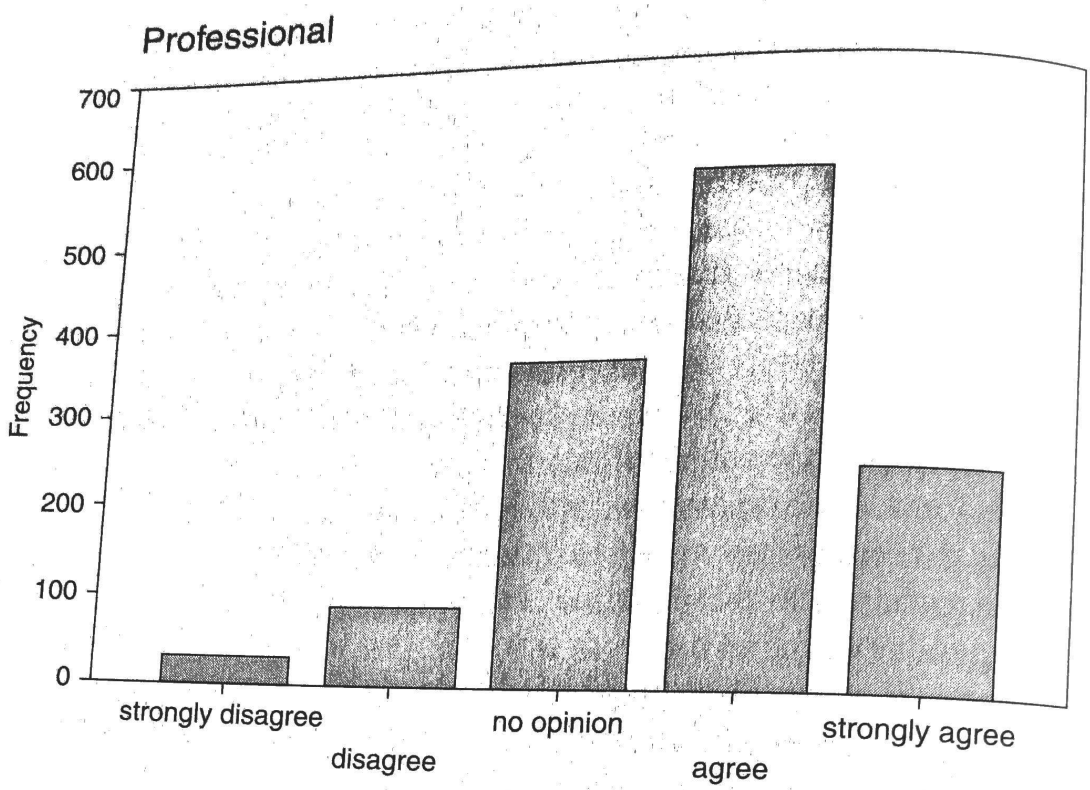

Figure 11-4 Bar Chart.

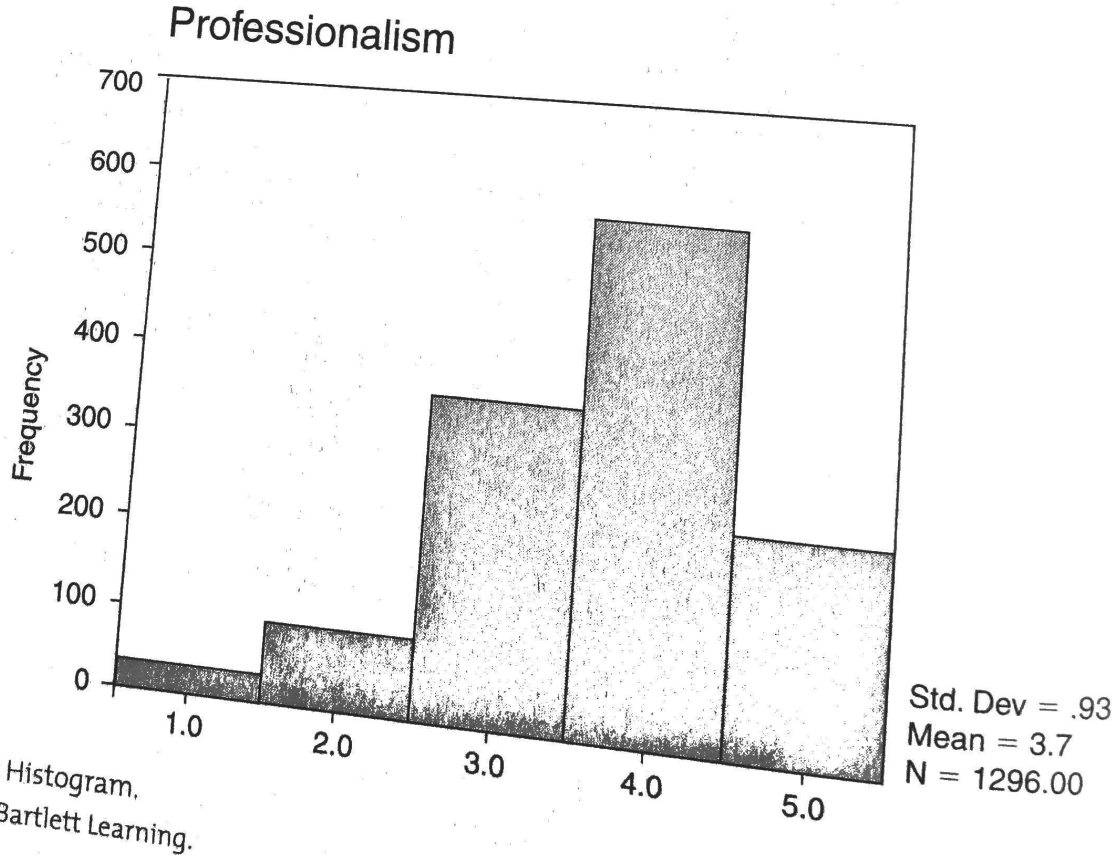

Figure 11-5 Histogram.

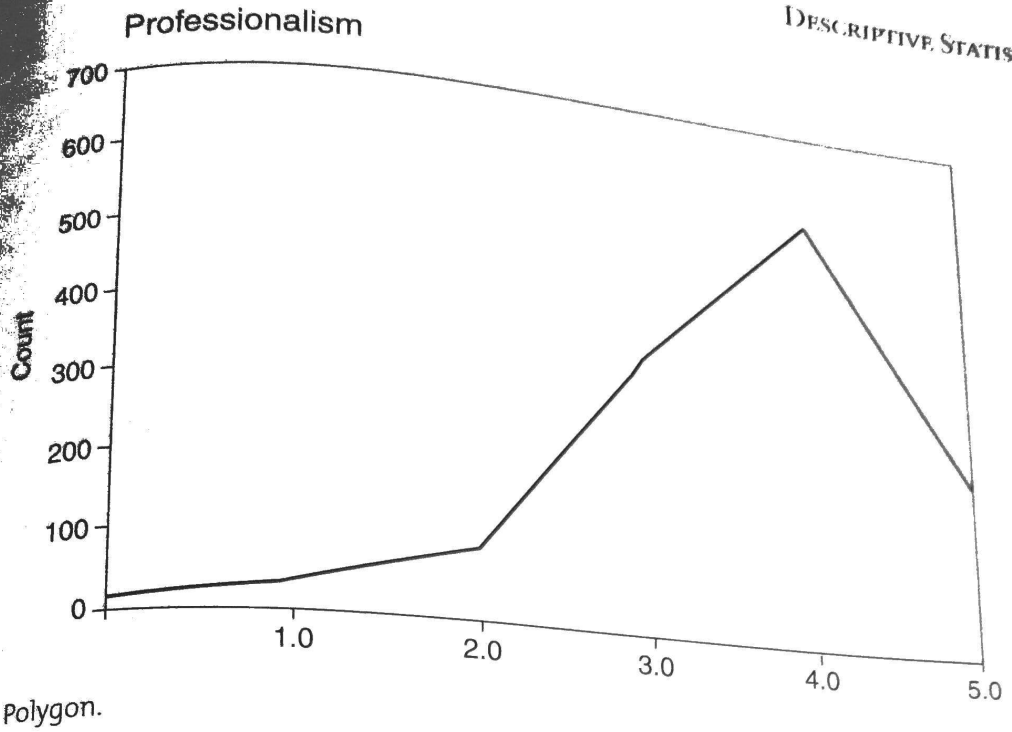

Figure 11-6 Polygon.

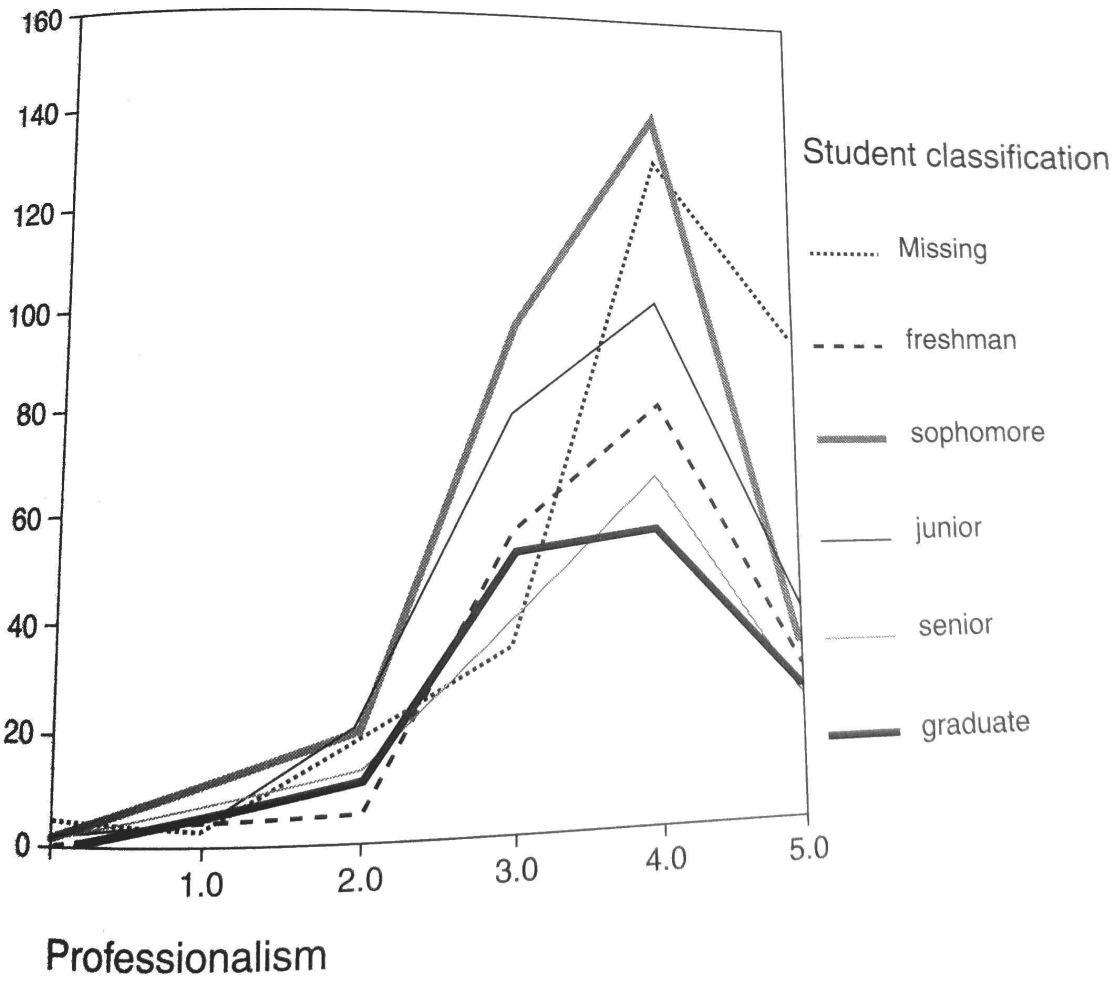

Figure 11-7 Line Graph.

one way to summarize data. The three most common measures are the (1) mean, (2) median, and (3) mode. The mean is the arithmetic average. The median is the midpoint or the number that falls in the middle. The mode is the number that occurs most frequently. The mean is usually used as a measure of central tendency for interval- or ratio-level data. The median is used mostly for ordinal-level data. The mode is generally used for nominal-level data. **Figure 11-8** shows how the three measures are obtained.

Using a measure of central tendency allows the researcher to simplify the numbers in a summary manner. In addition, these measures are often used in conjunction with measures of variability. Gaining a familiarity with the characteristics of these measures can help researchers identify when it might be appropriate to use measures of central tendency.

Measures of Variability

Despite their similarities, all statistics are not the same. Differences that occur among scores in a variable is called *variability*. The three main measures of variability are (1) range, (2) variance, and (3) standard deviation. The range is simply the difference between the highest and lowest scores. Variance is a summary measure of the differences between each of the scores and the mean. Standard deviation is also a summary measure of the differences between each of the scores and the mean, but it has been calculated to be the same unit of

Sample data n = 100

Years of College Completed	Number of Responses	Responses x Years of College
1	10	10
2	20	40
3	30	90
4	20	80
5	10	50
6	8	48
7	2	14
	100	332

Mean = average = 332/100 = 3.32 years of college
Mode = most frequent (30 responses) = 3 years of college
Median = midpoint (30 responses are below 3 years, 40 responses above 3 years, 30 responses are within the 3 years grouping. The midpoint would be between the 20th and 21st respondents in the 3-year group.)= 3 years of college.

Figure 11-8 Measures of Central Tendency.

measurement as the original variable. For example, if the original variable was age measured in years, then the standard deviation would also be measured in years, making it easier to interpret than the variance.

Skewness

The discussion thus far has alluded to the distribution of the data and has depicted it somewhat in sample figures. Many of the statistical techniques discussed in this chapter (and which will influence the discussions of inferential statistics in Chapter 12) are based on the assumption of a normal distribution. A normal distribution may also be referred to as a *normal*, or *bell*, *curve*. You may remember cutting hearts from paper as child for Valentine's Day. To do so, you folded the paper in half so that both sides of the heart would be identical or mirror images. If you were to cut a normal curve in half at the center point, which is the mean, then the sides would be mirror images of each other. By using polygons, line charts, or scatterplots (discussed in Chapter 12), researchers are able to visualize the distribution of the data. If it is a normal distribution, the researcher is able to use a broader range of statistical techniques. If it is a non-normal (also known as *nonparametric*) distribution, the statistical techniques that may be used are more limited.

The measures discussed previously (central tendency and variability) indicate the location of the center of the curve (central tendency) and the spread of the curve (variability). They may also differ based on the symmetry of the curve. If one side has more values so as to cause the slope of that side to stretch further outward and no longer create a mirror image, the distribution is said to be *skewed*. Skewness alerts the researcher to the presence of outliers, which are cases that have extremely high or low values. For example, if a line chart based on net worth included individuals such as Bill Gates and Warren Buffett, their fortunes would skew the data distribution. Gates and Buffett are the outliers in this data and pull the distribution to one side. Such knowledge aids the statistician in conducting analysis of the data.

Kurtosis

Kurtosis is another distribution consideration that warrants some discussion. *Kurtosis* refers to the amount of smoothness or pointedness of the curve. A tall, thin curve is described as being *leptokurtic*. A short, flat curve is *platykurtic*. For the purposes of this text, one need only recognize the terms if a statistician states that the data distribution exhibits such features.

These statistics, in tandem with measures of central tendency and variability, provide useful ways to describe the data. In general, descriptive statistics are primarily used to describe how the data are comprised. However, describing the data is usually just a small portion of making inferences from it.

Summary

Once the data collection is accomplished, the next two phases in completing the research are data preparation and data analysis. Because the data are often obtained in a raw manner, they must be coded for entry into a statistical software program. After the data have been entered, they must be cleaned to ensure accuracy and appropriateness for subsequent statistical analysis. At this point, missing data are dealt with, and recoding may also be required.

Univariate analysis using descriptive statistics is typically conducted to gain knowledge about the data before any bivariate and multivariate analyses are completed. Descriptive statistics may involve the use of frequency distribution tables and measures of central tendency and variability. Inferential statistics used in bivariate and multivariate analyses are discussed in more detail in Chapter 12.

APPLICATION EXERCISES

1. In Chapter 10, you described different data collection strategies for a study of workplace and school bullying. Of the ideas you presented to the directors of the violence prevention program, they approved data collection through the use of a mail survey. After receiving thousands of responses, you now need to analyze the data. The program directors have asked for results from the raw data. Explain the importance of first preparing the data for analysis.
2. One of the questions on the survey asked respondents how often they experience bullying by selecting a specific response from the following: did not experience bullying, once a month, 2–3 times per month, 4–5 times per month, and more than 5 times per month. Describe how you would code this information into a quantitative variable and identify the level of measurement.
3. Describe the process you would use to enter the data into a dataset.
4. Describe the process for cleaning the data.
5. Describe how you would address the issue of missing data.
6. Describe how you would recode the variable from step 2 into a nominal-level variable that indicates the respondent's experience with bullying and has the possible response categories of yes and no.
7. Describe the possible uses of univariate, bivariate, and multivariate analyses within the study.
8. Draw an example of each of the following methods for displaying frequencies:
   a. Pie chart
   b. Bar chart
   c. Histogram
   d. Line chart