Describing Data

DESCRIPTIVE STATISTICS AND EXPLORATORY DATA ANALYSIS

Descriptive statistics allow you to summarize the properties of an entire distribution of scores with just a few numbers. Although descriptive statistics are commonly used to address the specific questions that you had in mind when you designed your study, they also can be employed to help you discover important but perhaps hidden patterns in your data that may shed additional light on the problems you are interested in resolving. The search for such patterns in your data is termed **exploratory data analysis (EDA)**. Over the past 40 years or so, a whole new set of descriptive tools has been developed to aid you in this search, many of which are graphical in nature.

When research has been designed to answer a specific question or set of questions, there is a strong temptation to rush directly to the inferential statistical techniques that will assess the "statistical significance" of the findings and to request only those descriptive statistics related directly to the analysis, such as group means, standard deviations, and standard errors. Resist this temptation. As we explain in Chapter 14, many of the most commonly used inferential statistics make certain crucial assumptions about the populations from which the scores in your data set were drawn. If these assumptions are violated, the results of the statistical analysis may be misleading.

Some exploratory techniques help you spot serious defects in your data that may warrant taking corrective action before you proceed to the inferential analysis. Others help you determine which summary statistics would be appropriate for a given set of data. Still others may reveal unsuspected influences. In this chapter, we introduce you to a number of descriptive tools (both numerical and graphical) for describing data and revealing secrets that hide within.

ORGANIZING YOUR DATA

Before you can interpret your data, you must first organize and summarize them. How you organize your data depends on your research design (whether you have conducted a survey, observational study, or experiment), how many variables were observed and recorded, and how observations were grouped or subdivided. A few representative examples follow.

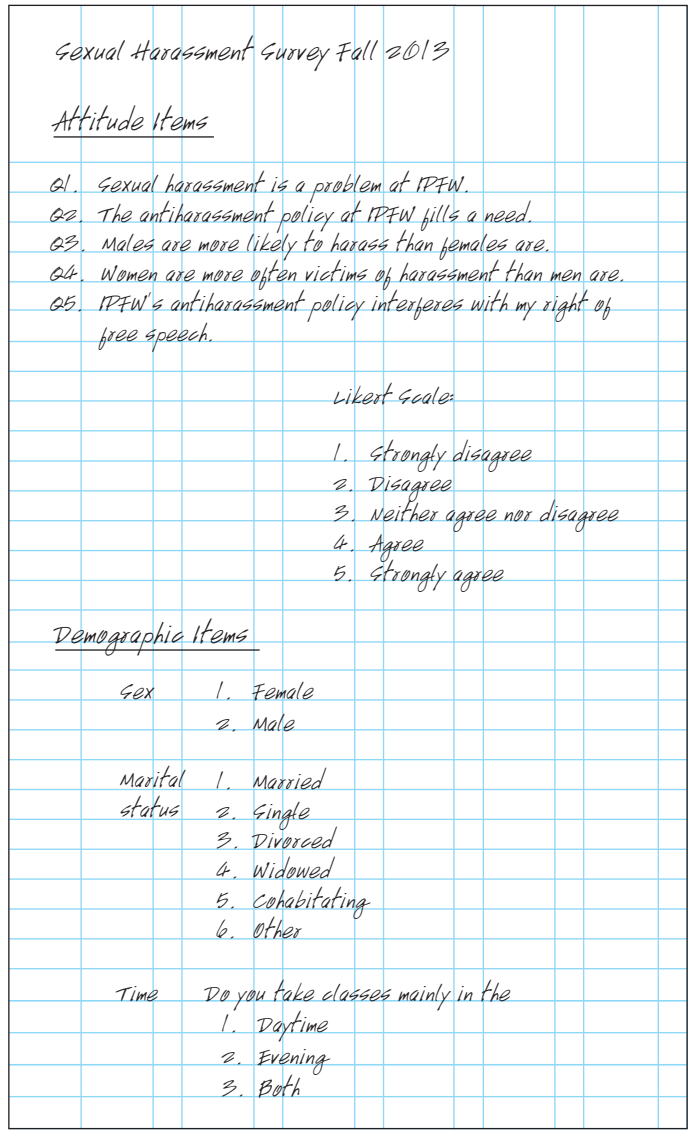
For *survey data,* a data summary sheet like that shown in Figure 13-1 would be appropriate. The data are organized into a series of columns, one for numbering the respondents, one for each question asked, and one for each demographic item. To save space, the identifiers over the columns should be kept short. Here, we have simply labeled each question Q1, Q2, and so on. If space permits, you may invent more descriptive labels. Each row gives the data for one respondent. In the example, certain demographic variables have been dummy-coded. **Dummy codes** identify category values as numbers (e.g., sex of respondent: 1 = female, 2 = male). By using dummy codes, you simplify data entry (e.g., you enter a "1" instead of "female").

Data sheets like the one shown in Figure 13-1 can go on for pages. A good strategy is to lay out your first data sheet and, before entering any data, copy it. In this way you avoid having to enter the column and row labels by hand on each new page.

Resp.	Q1	Q2	Q3	Q4	Q5	Sex	Age	Marit.	Time
1	3	4	4	4	2	2	25	2	1
2	4	3	5	5	2	1	20	2	3
3	3	3	4	4	3	1	*	1	1
4	1	4	4	4	2	2	22	2	1
5	3	4	4	4	2	2	29	1	3
6	3	3	4	4	3	1	19	2	1
7	2	4	3	4	1	2	19	2	1
8	2	4	4	5	4	2	25	2	3
9	2	3	3	3	3	1	21	2	1
10	2	3	2	3	3	2	20	2	1
11	4	3	4	4	3	2	18	2	3
33	2	4	3	4	2	1	24	1	1
34	2	3	2	4	3	1	24	2	1
35	1	2	4	4	4	2	31	3	2
36	2	4	5	5	3	1	26	2	3

FIGURE 13-1 Example data summary sheet for survey data (* = missing data).

In addition, you should make out a single "key" or "code sheet" that describes the scale(s) used for the questions, gives a fuller description of the question or variable in each column, and indicates what each dummy code represents. Figure 13-2 shows the code sheet that accompanies the data sheet shown in Figure 13-1. The code sheet shows the Likert-scale codes used in conjunction with each attitude item, the five attitude statements, the three demographic items, and the dummy codes used to represent the category values of sex, marital status, and time of class attendance.



Sexual Harassment Survey Fall 2013

Attitude Items

Q1. Sexual harassment is a problem at IPFW.
Q2. The antiharassment policy at IPFW fills a need.
Q3. Males are more likely to harass than females are.
Q4. Women are more often victims of harassment than men are.
Q5. IPFW's antiharassment policy interferes with my right of free speech.

Likert Scale:

1. Strongly disagree
2. Disagree
3. Neither agree nor disagree
4. Agree
5. Strongly agree

Demographic Items

Sex
1. Female
2. Male

Marital status
1. Married
2. Single
3. Divorced
4. Widowed
5. Cohabitating
6. Other

Time
Do you take classes mainly in the
1. Daytime
2. Evening
3. Both

FIGURE 13-2 Code sheet for the data summary sheet of Figure 13-1.

Subject number	Words/3 sec		Words/18 sec		CCCs/3 sec		CCCs/18 sec
1	19		19		16		16
2	18		17		16		06
3	19		14		20		14
4	19		16		15		14
5	19		15		19		13
6	19		17		13		07
7	20		20		18		13
8	19		19		14		06
9	20		19		17		14
10	18		13		05		02
11	19		16		14		11
12	18		17		08		02
13	18		13		11		00
14	16		06		11		03
31	19		19		14		08
32	20		17		19		17
33	19		14		07		00
34	20		14		16		10

FIGURE 13-3 Data summary sheet for a 2 × 2 within-subjects factorial experiment presented in an unstacked format.

Subject number	Item type (1 = word, 2 = ccc)		Retention interval (1 = 3s, 2 = 18s)		Number correct
1	1		1		19
1	1		2		19
1	2		1		16
1	2		2		16
2	1		1		18
2	1		2		17
2	2		1		16
2	2		2		06
3	1		1		19
3	1		2		14
3	2		1		20
3	2		2		14
34	1		1		20
34	1		2		14
34	2		1		16
34	2		2		10

FIGURE 13-4 Data summary sheet for the same data as shown in Figure 13-3, presented in stacked format.

Experimental or quasi-experimental designs break down the dependent variable according to treatments or categories. You can organize data from these designs in two distinct ways. One way (called an *unstacked format*) is to create a separate column for the scores from each treatment. Figure 13-3 shows a simple summary sheet organized in this way for a 2 × 2 within-subjects factorial experiment. The subject numbers appear in the leftmost column. Because each subject was exposed to all the treatments, only one column of subject numbers was needed. Reserve space at the bottom of the data summary sheet for column summary statistics such as the mean and standard deviation. You can enter these after you have analyzed the data.

The second way to organize your data is to use a *stacked format*. In this format, you create one column for the participant IDs, a column for the treatment levels (dummy-coded), and a column for each dependent variable. Figure 13-4 redisplays a portion of the data of Figure 13-3 in this way. The stacked format works better than the unstacked format when your data include multiple independent or dependent variables. These are easily accommodated by including

additional columns to indicate the treatment levels or observed values of the additional variables. Also, many computer statistical analysis packages expect the data to be entered in this format. A disadvantage of the stacked format is that, unlike the unstacked format, it does not provide a simple way to display treatment summary statistics.

More complex designs involving several independent and/or quasi-independent variables can be accommodated within either format. Figure 13-5 displays data in unstacked format from individual subjects for a 2 × 4 between-subjects design in a study by Bordens and Horowitz (1986) on the effects of joining multiple criminal offenses in a single trial (a procedure known as "joinder of offenses"). Each column provides the data from one treatment. In this two-factor between-subjects design, each treatment represented one level of "Charges judged" (the first independent variable) combined with one of the four levels of "Charges filed" (the second independent variable). The bottom two rows display summary measures (the mean and standard deviation) for each treatment.

A useful data summary sheet must be clearly labeled. Note that the columns of data in Figure 13-5 are clearly labeled with the levels of the independent variable in effect for each group. The top headings indicate the two levels of charges judged. The second level of headings indicates the level of charges filed as appropriate to each group.

	Charges judged							
	One charge				Two charges			
Charges filed	1	2	3	4	1	2	3	4
	3	5	6	2	6	3	4	5
	3	4	3	4	4	5	6	5
	3	4	4	5	4	5	4	5
	4	4	4	5	5	5	5	6
	4	5	5	5	4	4	5	5
	3	3	5	5	5	5	4	4
	2	3	4	5	5	4	4	5
	5	4	5	4	3	5	5	5
	6	4	3	6	3	5	5	6
	5	5	4	6	5	3	6	6
Mean	3.8	4.1	4.3	4.7	4.4	4.4	4.8	5.2
Standard deviation	1.23	0.74	0.95	1.16	0.97	0.84	0.79	0.63

FIGURE 13-5 Data summary sheet for a 2 × 4 between-subjects design.
SOURCE: Bordens and Horowitz, 1986.

The organization just described works well for a 2 × 4 factorial experiment and can be expanded to handle more levels of each factor or more factors. Other designs may require a different organization.

Organizing Your Data for Computer Entry

If you are going to submit your data to computer analysis, you should find out how the statistical analysis software that you intend to use expects the data to be organized. Many packages require the data to be entered in stacked format, some require the unstacked format, and some will accept either.

If you have not already done so on your original data sheets, you may have to code your variables before entering the data into the computer. Most software for data analysis looks for a numeric or alphabetic code to determine the levels or values of your independent and dependent variables. You must decide how to code these variables.

Coding independent variables involves assigning values to corresponding levels. For a quantitative independent variable (e.g., number of milligrams of a drug), simply record on your coding sheet the number of milligrams administered to subjects in each treatment group (e.g., 10, 20, or 30). For qualitative independent variables, you must assign an arbitrary number to each level. For example, if your independent variable were the loudness of a tone in an auditory discrimination experiment (low, moderate, and high), you might code the levels as 1 = low, 2 = moderate, 3 = high. As noted, this assignment of numbers to the levels of a qualitative independent variable is called *dummy coding*.

For quantitative data (e.g., if your participants rated the intensity of a sound on a scale ranging from 0 to 10), simply transfer each participant's score to your coding sheet. If, however, your dependent measure were qualitative (e.g., yes/no), you must dummy-code your dependent variable. For example, you could code all yes responses as 1 and all no responses as 2.

When coding your dependent variables, transfer the data (numeric or dummy-coded) to your coding sheet exactly as they are. Don't be concerned with creating new variables (e.g., by adding together existing ones) or with making special categories. Most statistical analysis software have commands that let you manipulate data in a variety of ways (adding numbers, doing data transformations such as a log transformation, etc.). So don't waste time creating new variables when preparing your data for input.

Entering Your Data

Personal computer versions of statistical software have easy-to-use spreadsheet interfaces that allow you to enter data quickly and make corrections easily. If you don't like the data-entry provision in a particular program, the program may allow you to enter your data into a stand-alone spreadsheet program such as Microsoft's Excel and then read the spreadsheet file into the statistical program's data editor.

You can make data entry easier by organizing your data-coding sheet the way your data editor expects the data to be entered. For example, if your data will be entered one column at a time, organize the data into columns on the sheet. Then simply read down the columns while entering your data.

Errors climb when fatigue sets in, so if you are entering a large amount of data, take frequent breaks. Be sure to save any data that you have entered before you leave the keyboard.

Speaking of saving data, nothing can be more frustrating than spending half an hour entering data, only to have them obliterated by an unexpected power failure. You can minimize your losses on such occasions by frequently saving your data. (Most programs have a backup feature that you can configure to save the data automatically at periodic intervals.) Also, don't forget to save your data before you turn off the computer or exit from the data editor. Any data that you fail to save will be lost. At this time, you also should make a backup copy of your data on another disk or some other device such as a USB drive.

When you save your data, you create a *data file* from which the data will be read when the computer conducts your analysis. When you save the file, the computer will ask you to type in a file name under which the data will be saved. Try to think of a descriptive file name that will uniquely identify the data. When you have several data files, using descriptive file names makes it easier to find the correct file. It is also a good idea to add a date to your file name so that you will know which is the most recent version (e.g., MemoryApril10.dat).

After you have entered your data, check for errors. Because the computer cannot detect incorrectly entered data, it is up to you to catch any mistakes if you are to avoid invalid results. If you have had someone else enter the data for you, don't assume that the other person has already done the checking.

Examining Your Data Typically, the data from individual groups or conditions are used to compute summary statistics, such as the group average, that represent group or condition as a whole. When you calculate an average, you have one score that characterizes an entire distribution. You then can refer to the performance of subjects in a group by citing the average performance. If your data will be submitted to a statistical analysis based on treatment means, you will be treating your data in this way.

Although convenient, group averages do have two important limitations. First, the average score may not represent the performance of individual subjects in a group. An average score of 5 can result if all 10 subjects in a group scored 5 or if half scored 0 and half scored 10. In the former case, the average accurately reflects the individual performance of each subject. In the latter case, it does not. We examine this idea in more detail during the discussion of the mean.

The second limitation of using grouped data is that a curve resulting from plotting averaged data may not reflect the true nature of the psychological phenomenon being studied. In a learning experiment, for example, in which rats must meet a learning criterion (e.g., three consecutive error-free trials), a graph showing how the group average changes across trials might suggest that learning is a gradual process. Inspection

of graphs of each individual subject's behavior might tell a different story. It might be that each rat evidences no learning for some variable number of trials, then suddenly masters the task, and after that never makes another error. Such a pattern of data would suggest that learning is an all-or-none proposition rather than the process of gradual improvement implied by the group average.

Unfortunately, researchers too often fall into the pattern of collecting data and then calculating an average without considering the individual scores constituting the average. A good strategy to adopt is to look at both the grouped and individual data. When you have repeated measures of the same behavior, examining individual data shows how each subject performed in your study. This may provide insights into the psychological process being studied that are not afforded by grouping data.

When you collect only a single score for each subject, you should still examine the distribution of individual scores. This usually entails plotting the individual scores on a graph and carefully inspecting the graph.

GRAPHING YOUR DATA

Whether you have chosen a grouped or an individual strategy for dealing with your data, you will often find it beneficial to plot your data on a graph. Graphing helps you make sense of your data by representing them visually. The next sections describe the various types of graphs and indicate their uses. For details on drawing graphs, see Chapter 16.

Elements of a Graph

A basic graph represents your data in a two-dimensional space. The two dimensions (horizontal and vertical) are defined by two lines intersecting at right angles, called the *axes* of the graph. The horizontal axis is called the *abscissa* or *x-axis* of the graph, and the vertical axis is called the *ordinate* or *y-axis*. (The terms *x-axis* and *y-axis* are used in this discussion.)

When graphing data from an experiment, you normally represent levels of your independent variable along the *x*-axis and values of the dependent variable along the *y*-axis. A pair of values (one for the *x*-axis and one for the *y*-axis) defines a single *point*

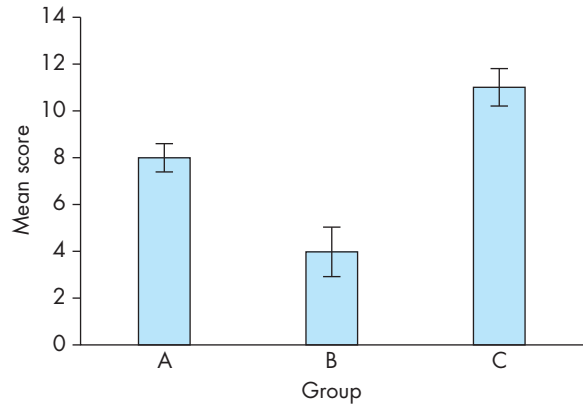


FIGURE 13-6 Bar graph from a hypothetical one-factor design, showing means and standard errors of the mean.

within the graph. You can present data within the two-dimensional space of a graph as a bar graph, line graph, scatter plot, or (abandoning the Cartesian *x*-axis, *y*-axis geometry) pie graph.

Bar Graphs

A **bar graph** presents your data as bars extending away from the axis representing your independent variable (usually the *x*-axis although this convention is not always followed). The length of each bar reflects the value of the dependent variable. Figure 13-6 shows group means from a one-factor, three-group experiment plotted as a bar graph. The three bars in Figure 13-6 represent the three levels of the independent variable for which data were collected. The length of each bar along the *y*-axis represents the mean score obtained on the dependent variable. Note that each bar straddles the *x*-axis value that it represents. The width of each bar has no meaning and is chosen to provide a pleasing appearance.

The bars usually represent estimates of population values based on sample data, such as the sample mean. In such cases the graph may also present an indication of the precision of the estimate in the form of *error bars,* whiskers that extend from the tops of the main bars. The error bars show the variability of scores around the estimate. Figure 13-6 displays error bars depicting the standard error of the mean.

You also can use a bar graph to represent data from a multifactor design. Figure 13-7 shows a bar graph of the data from the two-factor joinder of offenses experiment (Bordens & Horowitz, 1986) described previously. Notice that the four levels of number of charges filed (one to four) are placed along the *x*-axis. The two levels of charges judged (the second independent variable) are represented within the graph itself. The gray bars represent the data from the one-charge judged group whereas the colored bars represent the data from the two-charges judged group.

A bar graph is the best method of graphing when your independent variable is categorical (such as the type of drug administered). In this case, the distance along

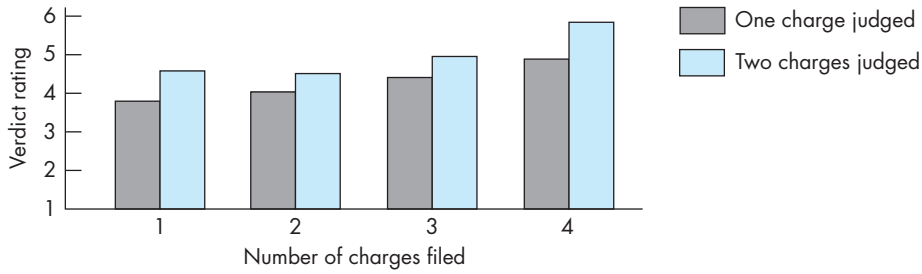


FIGURE 13-7 Bar graph of means from a two-factor design.

the *x*-axis has no real meaning. A line graph (which visually emphasizes that distance) would be misleading. The bar graph makes the arbitrary ordering of categories apparent, whereas a line graph would inappropriately suggest the presence of trends in these data.

In addition to displaying such statistical values as treatment means, bar graphs may be used to display certain kinds of data distributions, discussed later in the chapter.

Line Graphs

A **line graph** represents data as a series of points connected by a line. It is most appropriate when your independent variable, represented on the *x*-axis, is continuous and *quantitative* (e.g., the number of seconds elapsing between learning and recall). This is in contrast to a bar graph, which is most appropriate when your independent variable is categorical or *qualitative* (e.g., categories representing grades on an exam). Line graphs are also appropriate when you want to illustrate functional relationships among variables. A *functional relationship* is one in which the value of the dependent variable varies as a function of the value of the independent variable. Usually, the depicted functional relationship is causal.

Figure 13-8 illustrates a line graph that depicts the group means from a single-factor experiment with a continuous independent variable. The error bars extending vertically in both directions depict the precision of the points as estimates of the

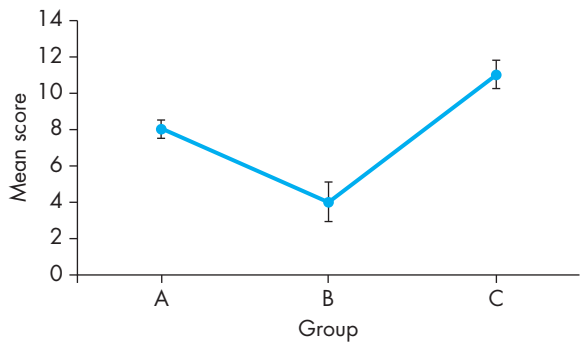


FIGURE 13-8 Line graph showing means and standard errors from a one-factor design.

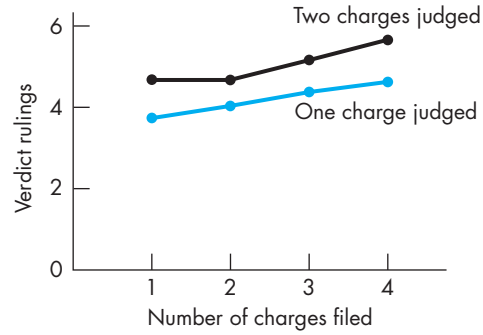


FIGURE 13-9 Line graph of means from a two-factor design.

population parameter, in this case represented by the standard error of the mean. These same data were shown in Figure 13-6 in the form of a bar graph. Notice the difference in how the two types of graphs visually represent the means.

A line graph also can be used to depict the means from multifactor experiments. Figure 13-9 shows such a line graph for the two-factor experiment on joinder of offenses (Bordens & Horowitz, 1986) described earlier. The levels of one factor are represented along the *x*-axis, just as in a single-factor experiment. The levels of the other factor are represented by using different symbols or line styles. All points collected under the same value of the second factor have the same symbol and are connected by the same line.

Shapes of Line Graphs Relationships depicted on a line graph can take a variety of shapes. Figure 13-10 shows a graph on which the curve is *positively accelerated.* A positively accelerated curve is relatively flat at first and becomes progressively steeper as it moves along the *x*-axis. Positive acceleration can occur both in the upward and downward directions along the *y*-axis.

A curve also may be *negatively accelerated,* as shown in Figure 13-11. Here the curve is steep at first but becomes progressively flatter as it moves along the *x*-axis. Eventually, the curve "levels off" at some maximum or minimum value. The function is said to be *asymptotic* at this value. The *asymptote* of a curve is its theoretical limit, or the point beyond which no further change in the value of the dependent variable is expected. In Figure 13-11 the relationship is asymptotic.

Whether positively or negatively accelerated, any curve also may be characterized as *increasing* or *decreasing,* which refers to whether the values along the *y*-axis

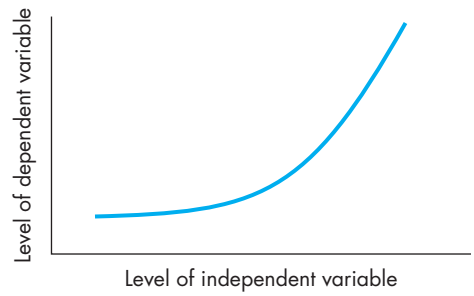


FIGURE 13-10 Line graph of positively accelerated functional relationship.

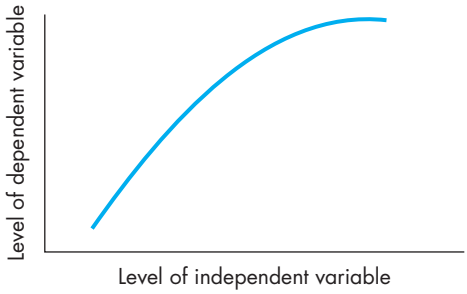


FIGURE 13-11 Line graph of negatively accelerated functional relationship.

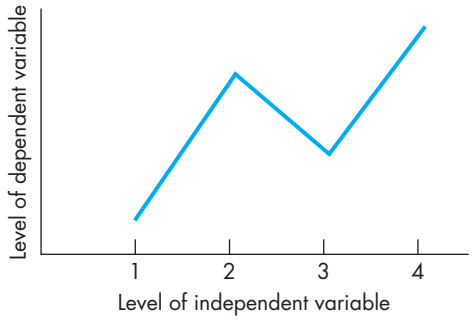


FIGURE 13-12 Line graph of nonmonotonic functional relationship.

increase or decrease, respectively, as the value along the *x*-axis increases. For example, a negatively accelerated, increasing function would approach a ceiling value at the asymptote whereas a negatively accelerated, decreasing function would approach a floor value.

A graph also may vary in complexity. The curves depicted in Figures 13-10 and 13-11 are both *monotonic.* That is, the curve represents a uniformly increasing or decreasing function. A *nonmonotonic* function contains reversals in direction, as illustrated in Figure 13-12. Notice how the curve changes direction twice by starting off low, rising, falling off, and then rising again.

Scatter Plots

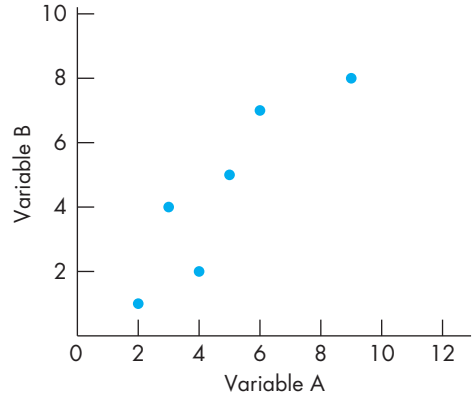
In research using a correlational strategy, the data from the two dependent measures are often plotted as a **scatter plot**. On a scatter plot, each pair of scores is represented as a point on the graph. For example, consider the data shown in Table 13-1. To make a scatter plot of these data, you plot the values of variable A along the *x*-axis and the values of variable B along the *y*-axis (or vice versa, it really does not matter). Then each pair of values is represented by a point within the graph. Figure 13-13 shows a scatter plot of the data in Table 13-1.

Scatter plots often include a "best-fitting" straight line (not shown in the figure) to indicate the general trend of the data points shown in the plot. In those cases the graph may also include the equation for this line and the coefficient of correlation. (We discuss these more fully in the section on correlation and regression.)

TABLE 13-1 Bivariate Data for a Scatterplot

SUBJECT NUMBER	VARIABLE A	VARIABLE B
1	5	5
2	4	2
3	9	8
4	2	1
5	6	7
6	3	4

FIGURE 13-13 Scatter plot of the bivariate data presented in Table 13-1.



Pie Graphs

If your data are in the form of proportions or percentages, then you might find a **pie graph** is a good way to represent the value of each category in the analysis. A pie graph represents the data as slices of a circular pie. Figure 13-14 shows two representative pie graphs. The pie graph to the left indicates the proportion of various behaviors observed in rat subjects during a half-hour coding period. The pie graph to the right, called an *exploded pie graph,* displays the same proportions while emphasizing the proportion of time devoted to grooming.

The Importance of Graphing Data

You can use either tables or graphs to summarize your data. If you organize data in tables, you present the numbers themselves (averages and/or raw score distributions). If you display the data in a graphical format, you lose some of this numerical precision. The value of a point usually can only be approximated by its position along the *y*-axis of the graph. However, graphing data is important for two major reasons, discussed in the next sections.

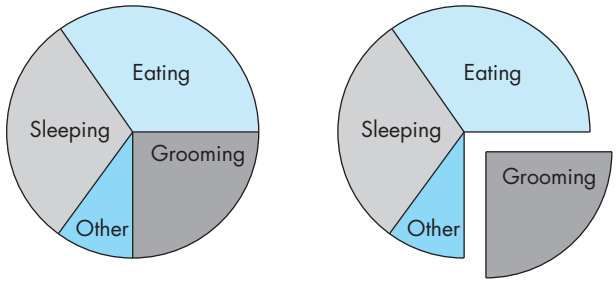


FIGURE 13-14 Pie graph and exploded pie graph.

Showing Relationships Clearly The saying "One picture is worth a thousand words" applies to graphing data from your research. Although summarizing data in a table is fine, proper graphing adds a degree of clarity no table can provide. The graph brings out subtleties in the relationships that may not be apparent from inspecting a table.

Choosing Appropriate Statistics In addition to making it easier to see relationships in your data, graphs allow you to evaluate your data for the application of an appropriate statistic. Before you apply any statistic to your data, graph your sample distributions and examine their shapes. Your choice of statistic will be affected by the manner in which scores are distributed, as described in the next section.

Graphing your data on a scatter plot is helpful when you intend to calculate a measure of correlation. Inspecting a scatter plot of your data can help you determine which measure of correlation is appropriate for your data. What you would look for and how your findings would affect your decision are taken up during the discussion of correlation measures later in the chapter.

THE FREQUENCY DISTRIBUTION

One of the first steps to perform when analyzing your data is to create a frequency distribution for each dependent variable in an experiment or for each variable in a correlational study. A **frequency distribution** consists of a set of mutually exclusive categories (*classes*) into which you sort the actual values observed in your data, together with a count of the number of data values falling into each category (*frequencies*). The classes may consist of response categories (e.g., for political party affiliation, they might consist of Democrat, Republican, Independent, and Other) or ranges of score values along a quantitative scale (e.g., for IQ, they might consist of 65–74, 75–84, 85–94, 95–104, 105–114, 115–124, and 125–134).

Displaying Distributions

Frequency distributions take the form of tables or graphs. Table 13-2 presents a hypothetical frequency distribution of IQ scores using the classes just given. Because IQ scores are quantitative data, the classes are presented in order of value from highest to lowest. To the right of each class is its frequency (f), the number of data values falling into that class. Because there were no IQ scores below 65 or above 134, classes beyond these limits are not tabled.

TABLE 13-2 Frequency Distribution Table of Hypothetical IQ Data

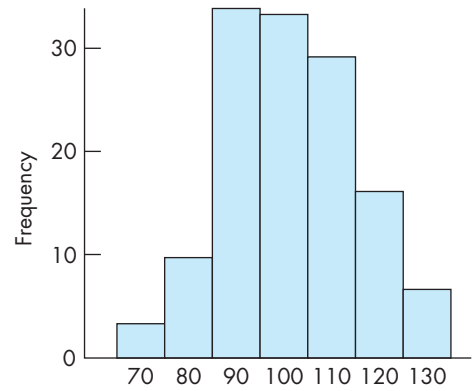
CLASS	f
125–134	5
115–124	12
105–114	22
95–104	25
85–94	26
75–84	7
65–74	3
Σf	100

Although a table provides a compact summary of the distribution, it is not particularly easy to extract useful information from it about center, spread, and shape. Graphical or semi-graphical displays are much better for this purpose. Here we describe two: the histogram and the stemplot.

The Histogram Figure 13-15 displays our IQ frequency distribution as a histogram. **Histograms** resemble bar graphs, with each bar representing a class. Unlike the bars in a bar graph, those in a histogram are drawn touching to indicate that there are no gaps between adjacent classes. Also, on a histogram, the *y*-axis represents a frequency: a count of the number of observations falling into a given category (e.g., the number of exam scores falling into the categories of A, B, C, D, or F). On a bar graph, the *y*-axis typically represents a mean score (e.g., the mean verdict rating shown in Figure 13-7).

The scale on which the variable was measured appears along the *x*-axis with the bars positioned appropriately to cover their respective ranges along the scale. The *y*-axis denotes the frequency; thus, a given bar's length indicates the frequency of scores falling within its range.

FIGURE 13-15 Hypothetical IQ data displayed as a histogram.



A histogram's appearance changes depending on how wide you make the classes. Make the classes too narrow, and you produce a flat-looking histogram with many empty or nearly empty classes. Make the classes too wide, and you produce a tall histogram lacking in detail. The goal is to create a histogram that shows reasonable detail without becoming flat and shapeless.

The Stemplot As a quick alternative to the histogram, you might consider using a **stemplot** (also known as a stem-and-leaf plot), which was invented by statistician John Tukey (1977), to simplify the job of displaying distributions. To create a stemplot of your data, you simply break each number into two parts: stem and leaf. The stem part might consist, for example, of the leftmost column or columns and the leaf part, the rightmost column. Thus, an IQ score of 67 would be broken into its leftmost number, or stem (6), and rightmost number, or leaf (7). After finding the lowest and highest stems, make a column that includes all the numbers in ascending order from lowest to highest stem. Then draw a vertical line immediately to the right of the stem column. Finally, for each score in your data, find its stem number and then write its leaf number on the same row immediately to the right of the stem. So, the IQ score of 67 would look like the first entry at the top of Figure 13-16. You do this for each number in your distribution. The final result would look something like Figure 13-16, which plots some hypothetical IQ data as a stemplot.

Stemplots are easy to construct and display and have the advantage over histograms and tables of preserving all the actual values present in the data. However, you do not have much freedom to choose the class widths because stemplots inherently create class widths of 10 (the span of a stem). Stemplots are not especially useful for larger data sets because the number of leaves becomes too large.

Examining Your Distribution

When examining a histogram or stemplot of your data, look for the following important features. First, locate the center of the distribution along the scale of measurement. In the IQ distribution plotted earlier, were the scores centered around 100 IQ points (an average value for the population as a whole) or somewhere else? The location of the center of a distribution tells you where the scores tended to cluster along the scale of measurement.

Second, note the spread of the scores. Do they tend to bunch up around the center or spread far from it? The spread of the scores indicates how variable they are.

Third, note the overall shape of the distribution. Is it hill shaped, with a single peak at the center, or does it have more than one peak? If hill shaped, is it more or less

Stem	Leaf
6	78
7	36
8	0222445555677789999
9	000001123444455667778889
10	011112222233445556666778899 9
11	0233334455568889
12	0012557
13	02

FIGURE 13-16 Hypothetical IQ data displayed as a stemplot.

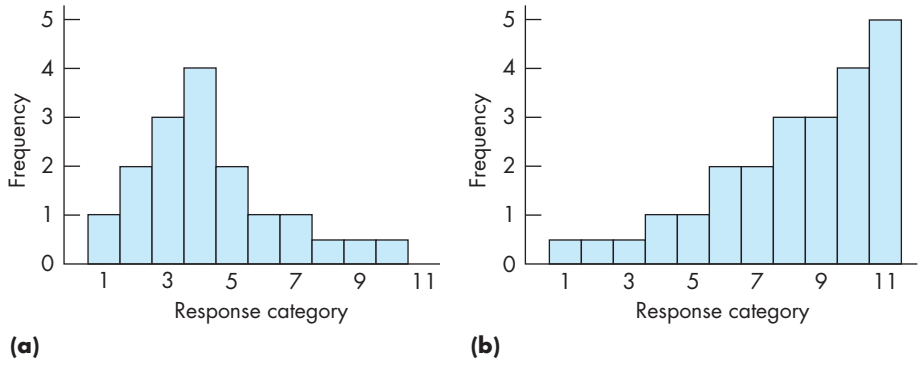


FIGURE 13-17 Two types of frequency distribution: (a) positively skewed and (b) negatively skewed.

symmetrical, or is it skewed? A **skewed distribution** has a long "tail" trailing off in one direction and a short tail extending in the other. A distribution is *positively skewed* if the long tail goes off to the right, upscale [see Figure 13-17(a)] or *negatively skewed* if the long tail goes off to the left, downscale [see Figure 13-17(b)]. Many variables encountered in psychology tend to produce a distribution that follows more or less a mathematical form known as the **normal distribution**, which is symmetric and hill shaped—the well-known bell curve. Because many common inferential statistics assume that the data follow a normal distribution, check the distribution of your data to see whether this assumption seems reasonable.

Finally, look for *gaps,* or **outliers**. Outliers are extreme scores that lie far from the others, well outside the overall pattern of the data (Moore & McCabe, 2006). Outliers may be perfectly valid (although unusual) scores, but sometimes they represent mistakes made in data collection or transcription. These bogus values can destroy the validity of your analysis. When you find an outlier, examine it carefully to determine whether it represents an error. Correct erroneous values or, if this is not possible, delete them from your analysis.

If you can find no valid reason for removing an outlier, you will have to live with it. However, you can minimize its effects on your analysis by using **resistant measures**, so called because they tend to resist distortion by outliers. We describe some of these measures next in our discussions of measures of center and spread.

DESCRIPTIVE STATISTICS: MEASURES OF CENTER AND SPREAD

In many research situations, it is convenient to summarize your data by applying descriptive statistics. This section reviews two categories of descriptive statistics: measures of center and measures of spread. The next section describes another category of descriptive statistics, measures of association.

Measures of Center

A **measure of center** (also known as a *measure of central tendency*) gives you a single score that represents the general magnitude of scores in a distribution. This score characterizes your distribution by providing information about the score at or near the middle of the distribution. The most common measures of center are the mode, the median, and the mean (also called the *arithmetic average*). Each measure of center has strengths and weaknesses. Also, situations exist in which a given measure of center cannot be used.

The Mode The **mode** is simply the most frequent score in a distribution. To obtain the mode, count the number of scores falling into each response category. The response category with the highest frequency is the mode. The mode of the distribution 1, 2, 4, 6, 4, 3, 4 is 4.

No mode exists for a distribution in which all the scores are different. Some distributions, called *bimodal distributions,* have two modes. Figure 13-18 shows a bimodal distribution.

Although the mode is simple to calculate, it is limited because the values of scores outside of the most frequent score are not represented. The only information yielded by the mode is the most frequent score. The values of other data in the distribution are not taken into account. Under most conditions, take into account the other scores to get an accurate characterization of your data. To illustrate this point, consider the following two distributions of scores: 2, 2, 6, 3, 7, 2, 2, 5, 3, 1 and 2, 2, 21, 43, 78, 22, 33, 72, 12, 8.

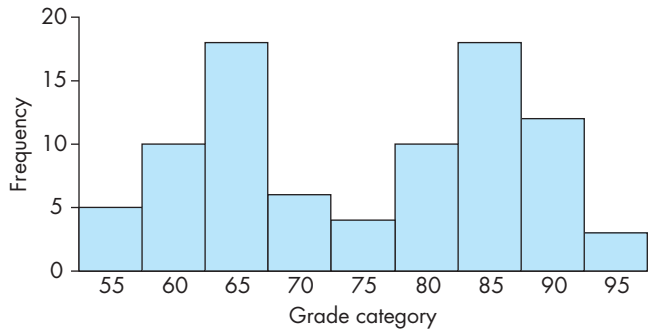


FIGURE 13-18 A bimodal frequency distribution.

In both these distributions, the mode is 2. Looking only at the mode, you might conclude that the two distributions are similar. Obviously, this conclusion is incorrect. It is clear that the second distribution is very different from the first. The mode may not represent a distribution very well and would not be the best measure to use when comparing distributions.

The Median A second measure of center is the median. The **median** is the middle score in an ordered distribution. To calculate the median, follow these steps:

1. Order the scores in your distribution from lowest to highest (or highest to lowest, it does not matter).
2. Count down through the distribution and find the score in the middle of the distribution. This score is the median of the distribution.

What is the median of the following distribution: 7, 5, 2, 9, 4, 8, 1? The correct answer is 5. The ordered distribution is 1, 2, 4, 5, 7, 8, 9, and 5 is the middle score.

You may be wondering what to do if you have an even number of scores in your distribution. In this case, there is no middle score. To calculate a median with an even number of scores, you order the distribution as before and then identify the *two* middle scores. The median is the average of these two scores. For example, with the ordered distribution of 1, 3, 6, 7, 8, 9, the median is 6.5 (6 + 7 = 13; 13/2 = 6.5).

The median takes more information into account than the mode. However, it is still a rather insensitive measure of center because it does not take into account the magnitudes of the scores above and below the median. As with the mode, two distributions can have the same median and yet be very different in character. For this reason, the median is used primarily when the mean is not a good choice.

The Mean The **mean** (denoted as M) is the most sensitive measure of center because it takes into account all scores in a distribution when it is calculated. It is also the most widely used measure of center. The computational formula for the mean is

$$M = \frac{\sum X}{n}$$

where $\sum X$ is the sum of the scores and n is the number of scores in the distribution. To obtain the mean, simply add together all the scores in the distribution and then divide by the total number of scores (n).

The major advantage of the mean is that, unlike the mode and the median, its value is directly affected by the magnitude of each score in the distribution. However, this sensitivity to individual score values also makes the mean susceptible to the influence of outliers. One or two such outliers may cause the mean to be artificially high or low. The following two distributions illustrate this point. Assume that distribution A contains the scores 4, 6, 3, 8, 9, 2, 3, and distribution B contains the scores 4, 6, 3, 8, 9, 2, 43. Although the two distributions differ by only a single score (3 versus 43), they differ greatly in their means (5 versus 10.7, respectively).

The mean of 5 appears to be more representative of the first distribution than the mean of 10.7 is of the second. The median is a better measure of center for the second distribution. The medians of the two distributions are 4 and 6, respectively—not nearly as different from one another as the means. Before you choose a measure of center, carefully evaluate your data for skewness and the presence of deviant, outlying scores. Do not blindly apply the mean just because it is the most sensitive measure of center.

Choosing a Measure of Center Which of the three measures of center you choose depends on two factors: the scale of measurement and the shape of the distribution of the scores. Before you use any measure of center, evaluate these two factors.

Chapter 5 described four measurement scales: nominal (qualitative categories), ordinal (rank orderings), interval (quantities measured from an arbitrary zero point), and ratio (quantities measured from a true zero point). The measurement scale that you chose when you designed your experiment will now influence your decision about which measure of center to use.

If your data were measured on a nominal scale, you are limited to using the mode. It makes no sense to calculate a median or mean sex, even if the sex of subjects has been coded as 0s (males) and 1s (females).

If your data were measured on an ordinal scale, you could properly use either the mode or the median, but it would be misleading to use the mean as your measure of center. This is because the mean is sensitive to the distance between scores. With an ordinal scale, the actual distance between points is unknown. You cannot assume that scores equally distant in terms of rank order are equally far apart, but you do assume this (in effect) if you use the mean.

The mean can be used if your data are scaled on an interval or ratio scale. On these two scales, the numerical distances between values are meaningful quantities.

Even if your dependent measure were scaled on an interval or ratio scale, the mean may be inappropriate. One of the first things you should do when summarizing your data is to generate a frequency distribution of the scores. Next, plot the frequency distribution as a histogram or stemplot and examine its shape. If your scores are normally distributed (or at least nearly normally distributed), then the mean, median, and mode will fall at the same point in the middle of the distribution, as shown in Figure 13-19. When your scores are normally distributed, use the mean as your measure of center because it is based on the most information.

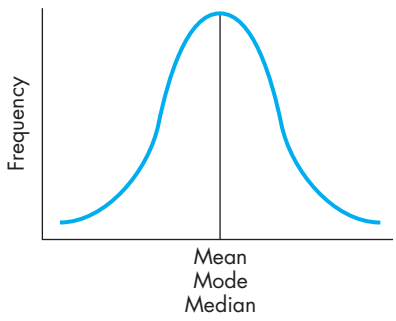


FIGURE 13-19 Line graph of normal distribution, showing location of mean, mode, and median.

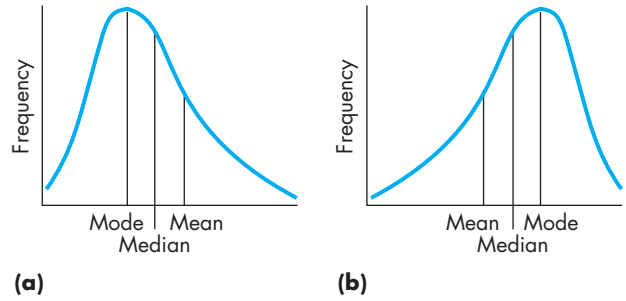


FIGURE 13-20 Line graph of (a) positively and (b) negatively skewed distributions, showing relationship between mean, mode, and median.

As your distribution deviates from normality, the mean becomes a less representative measure of center. The two graphs in Figure 13-20 show the relationship between the three measures of center with a positively skewed distribution and a negatively skewed distribution. Notice the relationship between the mean and median for these skewed distributions. In a negatively skewed distribution, the mean underestimates the center. Conversely, in a positively skewed distribution, the mean overestimates the center. Because the median is much less affected by skew, it provides a more representative picture of the distribution's center than does the mean and should be preferred whenever your distribution is strongly skewed.

Deviations from normality also create problems when deciding on an inferential statistic. Chapter 14 discusses inferential statistics and ways to deal with data that are not normally distributed. Neither the mean nor the median will accurately represent the center if your distribution is bimodal. With a bimodal distribution, both measures of center underrepresent one large cluster of scores and overrepresent the other.

Table 13-3 presents hypothetical scores from an introductory psychology exam that generated a bimodal distribution. These scores are shown graphically in Figure 13-18. The mean for these scores is 75.4, the median is 77, and both scores are in the grade C category. However, few students actually received a score in this range. The mean and median underrepresent the large cluster of scores in the grade B category and overestimate the large cluster of scores in the grade D category. Thus, neither the mean nor the median would be an appropriate measure of center for the scores in Table 13-3.

To summarize the discussion to this point, the three measures of center are the mean, the median, and the mode. The mean is the most sensitive measure of center because it takes into account the magnitude of each score in the distribution. The mean is also the preferred measure of center. The median is less sensitive to the distribution of scores than the mean but is preferred when your distribution is skewed or the distribution contains serious outliers. Which measure of center you can legitimately use depends on the scale on which the dependent variable was measured and on the manner in which the scores are distributed.

TABLE 13-3 Hypothetical Scores on an Exam in an Introductory Psychology Class

54	63	69	82	87
56	64	69	82	87
56	64	69	83	87
56	64	69	83	88
57	65	72	84	88
58	65	75	84	88
59	65	75	84	89
61	65	75	85	89
61	65	76	85	89
62	66	78	86	89
62	66	78	86	90
62	66	79	87	90
62	66	80	87	91
62	66	81	87	92
62	67	81	87	92
63	67	81	87	93
63	68	82	87	94

Measures of Spread

Another important descriptive statistic you should apply to your data is a **measure of spread** (also known as a *measure of variability*). If you look again at some of the sample distributions described thus far (or at the data presented in Table 13-1), you will notice that the scores in the distributions differ from each other. When you conduct an experiment, it is extremely unlikely that your subjects will all produce the same score on your dependent measure. A measure of spread provides information that helps you to interpret your data. Two sets of scores may have highly similar means yet very different distributions, as the following example illustrates.

Imagine that you are a scout for a professional baseball team and are considering one of two players for your team. Each player has a .263 batting average over 4 years of college. The distributions of the two players' averages are as follows:

Player 1: .260, .397, .200, .195

Player 2: .263, .267, .259, .263

Which of these two players would you prefer to have on your team? Most likely, you would pick player 2 because he is more "consistent" than player 1. This simple example

illustrates an important point about descriptive statistics. When you are evaluating your data, you should take into account both the center *and* the spread of the scores. This section reviews four measures of spread: the range, the interquartile range, the variance, and the standard deviation.

The Range The **range** is the simplest and least informative measure of spread. To calculate the range, you simply subtract the lowest score from the highest score. In the baseball example, the range for player 1 is .202, and the range for player 2 is .008.

Two problems with the range are that it does not take into account the magnitude of the scores between the extremes and that it is very sensitive to outliers in the distribution. Compare the following two distributions of scores: 1, 2, 3, 4, 5, 6 and 1, 2, 3, 4, 5, 31. The range for the first distribution is 5, and the range for the second is 30. The two ranges are highly discrepant despite the fact that the two distributions are nearly identical. For these reasons, the range is rarely used as a measure of spread.

The Interquartile Range The **interquartile range** is another measure of spread that is easy to calculate. To obtain the interquartile range, follow these steps:

1. Order the scores in your distribution.
2. Divide the distribution into four equal parts (quarters).
3. Find the score separating the lower 25% of the distribution (quartile 1, or Q_1) and the score separating the top 25% from the rest of the distribution (Q_3). The interquartile range is equal to Q_3 minus Q_1.

The interquartile range is less sensitive than the range to the effects of extreme scores. It also takes into account more information because more than just the highest and lowest scores are used for its calculation. The interquartile range may be preferred over the range in situations in which you want a relatively simple, rough measure of spread that is resistant to the effects of skew and outliers.

The Variance The **variance** (s^2) is the average squared deviation from the mean. The defining formula is

$$s^2 = \frac{\sum(X - M)^2}{n - 1}$$

where X is each individual score making up the distribution, M is the mean of the distribution, and n is the number of scores. Table 13-4 shows how to use this formula by means of an example worked out for one distribution of scores.

The Standard Deviation Although the variance is frequently used as a measure of spread in certain statistical calculations, it does have the disadvantage of being expressed in units different from those of the summarized data. However, the variance can be easily converted into a measure of spread expressed in the *same* unit of measurement as the original scores: the **standard deviation** (s). To convert from the variance to the standard deviation, simply take the square root of the variance. The standard deviation of the data in Table 13-4 is 2.61 $\left(\sqrt{6.8}\right)$. The standard deviation is the most popular measure of spread.

TABLE 13-4 Calculation of a Variance

	X	X^2	$(X-M)$	$(X-M)^2$
	3	9	−2	4
	5	25	0	0
	2	4	−3	9
	7	49	2	4
	9	81	4	16
	4	16	−1	1
Σ	30	184		34

$M = 30/6 = 5.0$

$s^2 = 34/5 = 6.8$

MEASURES OF ASSOCIATION, REGRESSION, AND RELATED TOPICS

In some cases, you may want to evaluate the direction and degree of relationship (correlation) between the scores in two distributions. For this purpose, you must use a *measure of association*. This section discusses several measures of association, along with the related topics of linear regression, the correlation matrix, and the coefficient of determination.

The Pearson Product-Moment Correlation Coefficient

The most widely used measure of association is the **Pearson product-moment correlation coefficient, or Pearson *r***. You would use it when you scale your dependent measures on an interval or a ratio scale. The Pearson correlation coefficient provides an index of the direction and magnitude of the relationship between two sets of scores.

The value of the Pearson *r* can range from +1 through 0 to −1. The sign of the coefficient tells you the direction of the relationship. A positive correlation indicates a *direct relationship* (as the values of the scores in one distribution increase, so do the values in the second). A negative correlation indicates an *inverse relationship* (as the value of one score increases, the value of the second decreases). Figure 13-23 illustrates scatter plots of data showing positive, negative, and no correlation.

The magnitude of the correlation coefficient tells you the degree of *linear relationship* (straight line) between your two variables. A correlation of 0 indicates that no relationship exists. As the strength of the relationship increases, the value of the correlation coefficient increases toward either +1 or −1. Both +1 and −1 indicate a perfect linear relationship. The sign is unrelated to the magnitude of the relationship and simply indicates the direction of the relationship. Figure 13-24 shows three correlations of differing strengths. Panel (a) shows a correlation of +1; panel (b), a correlation of about +.8; and panel (c), a correlation of 0.

Factors That Affect the Pearson r Before you use the Pearson *r*, examine your data much as you do when deciding on a measure of center. Several factors affect the magnitude and sign of the Pearson *r*.

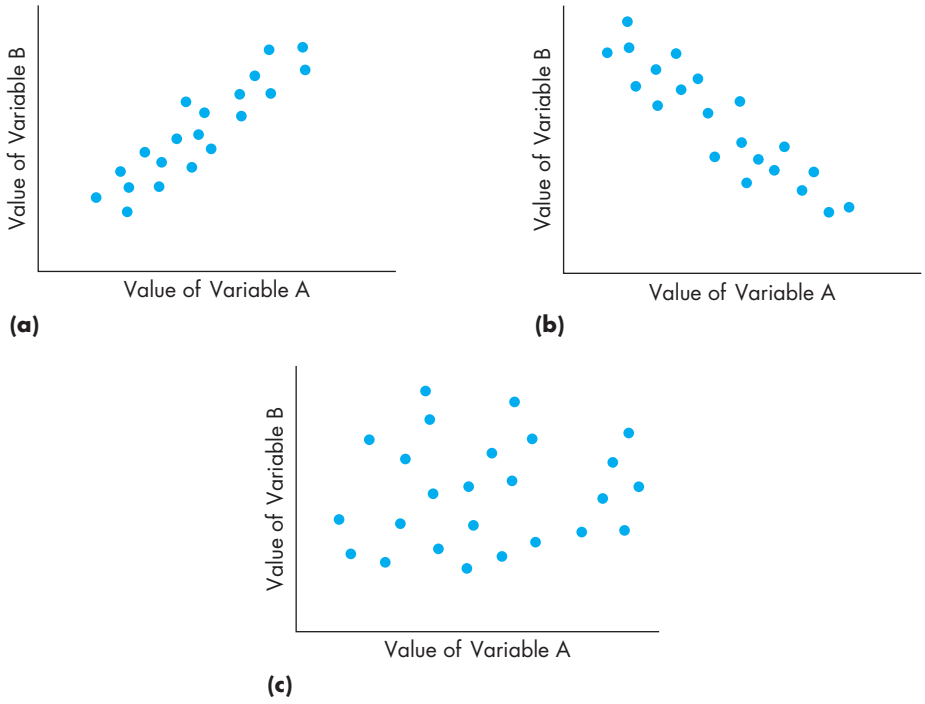


FIGURE 13-23 Scatter plots showing (a) positive, (b) negative, and (c) no correlation.

The presence of outliers is one factor that affects the Pearson *r*. An outlier can drastically change your correlation coefficient and affect the magnitude of your correlation, its sign, or both. This is especially true if you use a small number of pairs of scores to compute the Pearson *r*.

Restricting the range over which the variables vary also can affect Pearson *r*. For example, if you were to examine the relationship between IQ and grade point average (GPA) in a group of college students, you would probably find a weaker correlation than if you examined the same two variables using high school students. Because IQ varies less among college students than among high school students, any variation in GPA that relates to IQ also will tend to vary less. As a result, the impact of extraneous variables such as motivation will be relatively larger, leading to a reduced correlation.

The Pearson *r* is sensitive to not only the range of the scores but also the shapes of the score distributions. The formula used to calculate the coefficient uses the standard deviation for each set of scores. Recall that you use the mean to calculate the standard deviation. If the scores are not normally distributed, the mean does not represent the distribution well. Consequently, the standard deviations will not accurately reflect the variability of the distributions, and the correlation coefficient will not provide an accurate index of the relationship between your two sets of scores. Hence, you should inspect the frequency distributions of each set of scores to ensure that they are normal (or nearly normal) before using the Pearson *r*.

Finally, the *Pearson r* reflects the degree to which the relationship between two variables is linear. Because of this assumption, take steps to determine whether the relationship appears to be linear. You can do this by constructing a scatter plot and then

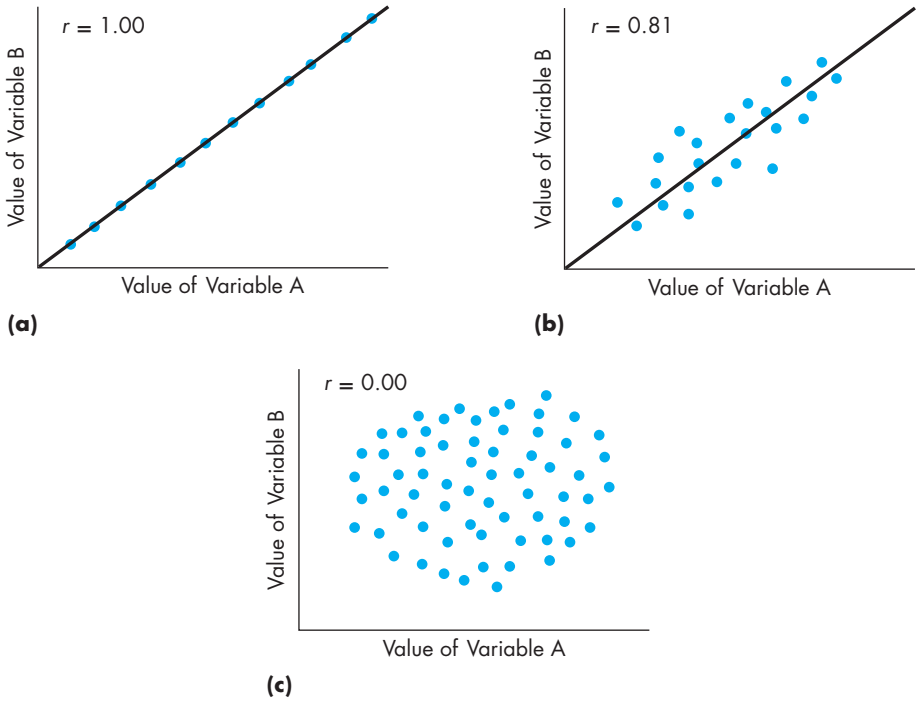


FIGURE 13-24 Scatter plots showing correlations of differing strengths: (a) perfect positive correlation, (b) strong positive correlation, and (c) zero correlation.

determining whether the points appear to scatter symmetrically around a straight line. Figure 13-25 shows a scatter plot in which the measures have a *curvilinear relationship* (rather than a linear relationship).

When the relationship between variables is nonlinear, the Pearson *r* underestimates the degree of relationship between the variables. For example, the Pearson correlation between the variables illustrated in Figure 13-25 is zero. However, the two variables are obviously systematically related. There are special correlation techniques for nonlinear data, which are not discussed here.

The Pearson *r* is used when both of your variables are measured along a continuous scale. You may need to correlate variables when one (or both) of them is not measured along a continuous scale. Special correlation coefficients are designed for these purposes, three of which are discussed in the next sections.

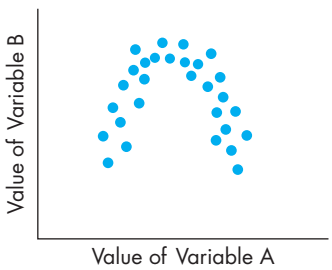


FIGURE 13-25 Scatter plot showing a curvilinear relationship.

The Point-Biserial Correlation

You may have one variable measured on an interval scale and the other measured on a nominal scale. For example, perhaps you want to investigate the relationship between self-rated political conservatism (measured on a 10-point scale) and whether or not a person voted for a particular referendum (yes or no). Because one variable is continuous and the other dichotomous (able to take on one of only two values), you would apply the **point-biserial correlation**.

Although there is a special formula for the point-biserial correlation, in practice you use the formula for the Pearson *r* to compute it. The dichotomous variable is dummy-coded as 0 for one response and 1 for the other. It is easier to use the Pearson formula, especially if you are using a computer program to evaluate your data (assuming the program cannot compute a point-biserial correlation).

Factors That Affect the Point-Biserial Correlation You should know a couple of things about the point-biserial correlation. First, its magnitude partly depends on the proportion of participants falling into each of the dichotomous categories. If the number of participants in each category is equal, then the maximum value the point-biserial can attain is ± 1.0 (just as with the Pearson *r*). However, if the number of participants in each category is *not* equal, then the maximum attainable value for the point-biserial correlation is less than ± 1.0. Consequently, the degree of relationship between the two variables may be underestimated. You should examine the proportion of participants using each category of the dichotomous variable and, if the proportions differ greatly, temper your conclusions accordingly.

The magnitude of the point-biserial correlation also is affected by the limited variation of the dichotomous variable (i.e., only two values possible). If the underlying variable is continuous but has been dichotomized for the analysis (e.g., anxiety level specified as either low or high), the point-biserial correlation will tend to underestimate the true strength of the relationship.

The Spearman Rank-Order Correlation

The **Spearman rank-order correlation**, or **rho** (ρ), is used either when your data are scaled on an ordinal scale (or greater) or when you want to determine whether the relationship between variables is monotonic (Gravetter & Wallnau, 2013). The rank-order correlation is relatively easy to calculate and can be interpreted in much the same way as a Pearson *r*.

The Phi Coefficient

The **phi coefficient (φ)** is used when *both* of the variables being correlated are measured on a dichotomous scale. You can calculate the phi coefficient with its own formula. However, like the point-biserial correlation, phi is usually calculated by dummy-coding the responses as 1s and 0s and then plugging the resulting scores into the formula for the Pearson *r*. The same arguments concerning restriction of range that apply to the point-biserial correlation also apply to phi—only doubly so.

QUESTIONS TO PONDER

1. What do measures of association tell you?
2. What are the measures of association available to you, and when would you use each?
3. What affects the magnitude and direction of a correlation coefficient?

Linear Regression and Prediction

A topic closely related to correlation is **linear regression**. With simple correlational techniques, you can establish the direction and degree of relationship between two variables. With linear regression, you can estimate values of a variable based on knowledge of the values of others. The following section introduces you to simple bivariate (two-variable) regression (also included are some calculations to help you understand regression). Chapter 15 extends bivariate regression to the case in which you want to consider multiple variables together in a single analysis.

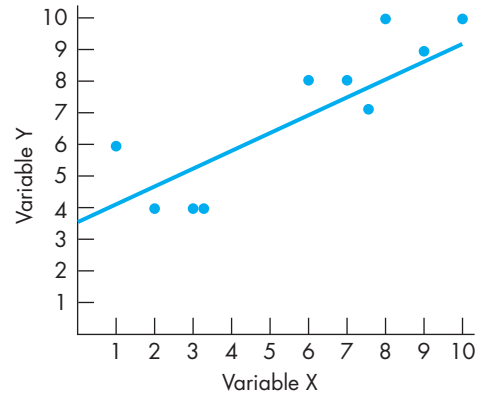
Bivariate Regression The idea behind **bivariate linear regression** is to find the straight line that best fits the data plotted on a scatter plot. Consider an example using the data presented in Table 13-5, which shows the scores for each of 10 subjects on two measures (X and Y). Figure 13-26 shows a scatter plot of these data. You want to find the straight line that best describes the linear relationship between X and Y.

The best-fitting straight line is the one that minimizes the sum of the squared distances between each data point and the line, as measured along the y-axis (least-squares

TABLE 13-5 Data for Linear Regression Example

X	Y	$(X-M)$	$(Y-\bar{Y})$	$(X-M)(Y-\bar{Y})$	$(X-M)^2$
7	8	1.40	1.30	1.82	1.96
3	4	−2.60	−2.70	7.02	6.76
2	4	−3.60	−2.70	9.72	12.96
10	9	4.40	2.30	10.12	19.36
8	9	2.40	2.30	5.52	5.76
7	7	1.40	0.30	0.42	1.96
9	8	3.40	1.30	4.42	11.56
6	8	0.40	1.30	0.52	0.16
3	4	−2.60	−2.70	7.02	6.76
1	6	−4.60	−0.70	3.22	21.16
$M_X = 5.6$	$M_Y = 6.7$			$SP = 49.8$	$SS_X = 88.4$

FIGURE 13-26 Scatter plot of data from Table 13-5.



criterion). This line is called the **least-squares regression line**. At any given value for *X* found in the data, the position of the line indicates the value of *Y* predicted from the linear relationship between *X* and *Y*. You can then compare these predicted values with the values actually obtained. The best-fitting straight line minimizes the squared differences between the predicted and obtained values.

Of course, you usually don't know if the relationship is truly a causal one or in which direction the causal arrow points. Consequently, you should interpret this statistic with caution. Perhaps the most enlightening use of this statistic is to subtract it from 1.0. The resulting number, called the **coefficient of nondetermination**, gives the proportion of variance in one variable *not accounted for* by variance in the other variable. This is in effect *unexplained* variance caused by unmeasured factors. If the coefficient of nondetermination is large, then your measured variables are having little impact on each other relative to these unmeasured factors. If this happens, then perhaps you should try to identify these unmeasured variables and either hold them constant or measure them.

The Correlation Matrix

If you have computed all the possible correlations among a number of variables, you can make the relationships among the variables easier to comprehend by displaying the correlation coefficients in a table called a **correlation matrix**. Table 13-6 shows a hypothetical correlation matrix for five variables (1 to 5). The variables being correlated in the matrix are shown in the headings along the top and left side of the matrix. Each number within the matrix is the correlation between the two variables whose row and column intersect at the position of the number. For example, the correlation between variables 5 and 3 can be found by reading across the row labeled "variable 5" to the column labeled "variable 3." The correlation found at that intersection is .06.

Note that the numbers along the diagonal are omitted from the table. This is because the diagonal positions represent the correlations of each variable with itself, which are necessarily 1.0. You also omit the correlations above the diagonal because they simply duplicate the correlations already given below the diagonal. For example,

TABLE 13-6 A Correlation Matrix

	VARIABLES			
VARIABLES	*1*	*2*	*3*	*4*
2	.54			
3	.43	.87		
4	.52	.31	.88	
5	.77	.44	.06	.39

the correlation of variable 5 with variable 3 (below the diagonal) is the same as the correlation of variable 3 with variable 5 (which would appear above the diagonal).

Multivariate Correlational Techniques

The measures of correlation and linear regression discussed in this chapter are all bivariate. Even if you calculate several bivariate correlations and arrange them in a matrix, your conclusions are limited to the relationship between pairs of variables. Bivariate correlation techniques are certainly useful and powerful tools. In many cases, however, you may want to look at three or more variables simultaneously. For example, you might want to know what the relationship between two variables is with the effect of a third held constant. Or you might want to know how a set of predictor variables relates to a criterion variable. In these cases and related others, the statistical technique of choice is *multivariate analysis.* Multivariate analysis is a family of statistical techniques that allow you to evaluate complex relationships among three or more variables. Multivariate analyses include multiple regression, discriminant analysis, part and partial correlation, and canonical correlation. Chapter 15 provides an overview of these and other multivariate techniques.

SUMMARY

When you have finished conducting your research, you begin the task of organizing, summarizing, and describing your data. The first step is to organize your data so that you can more easily conduct the relevant analyses. A good way to gain some understanding of your data is to graph the observed relationships. You can do this with a bar graph, line graph, scatter plot, or pie graph, whichever is most appropriate for your data.

A frequency distribution shows how the scores in your data vary along the scale of measurement. Although a frequency distribution can be presented in tabular format, you can grasp its essential features more easily by graphing it as a histogram or by creating a stemplot. When examining these, you should look for several important features: the center, around which the scores tend to vary; the spread, or degree to which the scores tend to vary from the center; the overall shape of the distribution (e.g., symmetric or skewed); and the presence of gaps or outliers—deviant points lying far from the rest. Examine any outliers carefully and correct or eliminate any that resulted from error. Descriptive statistics are methods for summarizing your data. Descriptive statistics include measures of central tendency, measures of variability, and measures of correlation.

The mode, median, and mean are the three measures of center. The mode is the most frequent score in your distribution. The median is the middle score in an ordered distribution. The mean is the arithmetic average of the scores, obtained by summing the scores and dividing the sum by the total number of scores.

Which of the three measures of center that you should use depends both on the scale that the data were measured on, and on the shape of the distribution of scores. The mean can be used only with data that are scaled on either a ratio or an interval scale and are normally distributed. In cases in which the data are skewed or bimodal, then the mean does not provide a representative measure of center, and the median or mode should be considered. Ordinally scaled data are best described with the median, and nominally scaled data are best described with the mode.

Measures of spread include the range, interquartile range, variance, and standard deviation. The range is simply the difference between the highest and lowest scores in your distribution. Although simple to calculate, the range is rarely used. Serious limitations of the range are that it is strongly affected by extreme scores and takes into account only the highest and lowest scores (thus ignoring the remaining scores in the distribution). The interquartile range takes into account more of the scores in the distribution and is less sensitive than the range to extreme scores. The variance uses all the scores in its calculation but has the disadvantage that its unit of measurement differs from that of the scores from which it derives. This problem can be overcome by taking the square root of the variance. The resulting statistic, the standard deviation, is the most commonly used measure of spread.

Your decision about which of the measures of spread to use is affected by the same two factors that affect your decision about central tendency (scale of measurement and distribution of scores). The standard deviation is a good measure of spread when your scores are normally distributed. As scores deviate from normality, the standard deviation becomes a less representative measure of spread. When your data are skewed, use the interquartile range.

The five-number summary provides a concise view of your distribution by providing the minimum, first quartile, median, third quartile, and maximum. Displaying these five numbers as a boxplot helps visualize the center, spread, and shape of the distribution. You can quickly compare distributions of the same variable from different treatments or samples by creating side-by-side boxplots.

Measures of association provide an index of the direction and degree of relationship between two variables. The most popular measure of association is the Pearson product-moment correlation coefficient (r). This coefficient can range from -1 through 0 to $+1$. A stronger relationship is indicated as the coefficient approaches ± 1. A negative correlation indicates that an increase in the value of one variable is associated with a decrease in the value of the second (inverse relationship). A positive correlation indicates that the two measures increase or decrease together (direct relationship).

The Pearson r is applied to data scaled on either an interval or a ratio scale. Other measures of correlation are available for data measured along other scales. The point-biserial correlation is used if one variable is measured on an interval or ratio scale and the other on a dichotomous nominal scale. Spearman's rho is used if both variables are measured on at least an ordinal scale. The phi coefficient is used if both variables are dichotomous.

Linear regression is a statistical procedure closely related to correlation. With linear regression, you can estimate the value of a criterion variable given the value of a predictor. In linear regression, you calculate a least-squares regression line, which is the straight

line that best fits the data on a scatter plot. This line minimizes the sum of the squared distances between each data point and the line, as measured along the *y*-axis (least-squares criterion), and minimizes the difference between predicted and obtained values of *y*. The amount of discrepancy between the values of *y* predicted with the regression equation and the actual values is provided by the standard error of estimate. The magnitude of the standard error is related to the magnitude of the correlation between your variables. The higher the correlation, the lower the standard error.

By squaring the correlation coefficient, you obtain the coefficient of determination, an index of the amount of variation in one variable that can be accounted for by variation in the other. Subtracting the coefficient of determination from 1.0 gives you the coefficient of nondetermination, the proportion of variance *not* shared by the two variables. The larger this number, the larger the effect of unmeasured sources of variance relative to that of the measured variables.

Multivariate statistical techniques are used to evaluate more complex relationships than simple bivariate statistics. With multivariate statistics, you can analyze the degree of relationship between a set of predictor variables and a criterion variable or look at the correlation between two variables with the effect of a third variable held constant.

KEY TERMS

descriptive statistics
exploratory data analysis (EDA)
dummy code
bar graph
line graph
scatter plot
pie graph
frequency distribution
histogram
stemplot
skewed distribution
normal distribution
outlier
resistant measure
measure of center
mode
median
mean
measure of spread
range
interquartile range
variance
standard deviation
five-number summary
boxplot
Pearson product-moment correlation coefficient, or Pearson *r*
point-biserial correlation
Spearman rank-order correlation (rho)
phi coefficient (φ)
linear regression
bivariate linear regression
least-squares regression line
regression weight
standard error of estimate
coefficient of determination
coefficient of nondetermination
correlation matrix