CHAPTER

2

Descriptive Statistics

Watch screencasts of the guided examples in this chapter. **edge.sagepub.com/pollock**

Procedures Covered

Analyze ▶ Descriptive Statistics ▶ Frequencies
Data ▶ Weight Cases
Graphs ▶ Legacy Dialogs ▶ Histogram
Analyze ▶ Reports ▶ Case Summaries

Analyzing descriptive statistics is the most basic—and sometimes the most informative—form of analysis you will do. Descriptive statistics reveal two attributes of a variable: its typical value (central tendency) and its spread (degree of dispersion or variation). The precision with which you can describe central tendency for any given variable depends on the variable's level of measurement. For nominal-level variables you can identify the *mode*, the most common value of the variable. For ordinal-level variables, those whose categories can be ranked, you can find the mode and the *median*—the value of the variable that divides the cases into two equal-size groups. For interval-level variables you can obtain the mode, median, and arithmetic *mean*, the sum of all values divided by the number of cases.

In this chapter you will use the Analyze ▶ Descriptive Statistics ▶ Frequencies procedure to obtain appropriate measures of central tendency, and you will learn to make informed judgments about variation. With the correct prompts, the Frequencies procedure also provides valuable graphic support—bar charts and (for interval variables) histograms. These tools are essential for distilling useful information from datasets having hundreds of anonymous cases, such as the American National Election Study (NES) or the General Social Survey (GSS). For smaller datasets with aggregated units, such as the States and World datasets, SPSS offers an additional procedure: Analyze ▶ Reports ▶ Case Summaries. Case Summaries lets you see firsthand how specific cases are distributed across a variable that you find especially interesting.

HOW SPSS STORES INFORMATION ABOUT VARIABLES

Suppose you were hired by a telephone-polling firm to interview a large number of respondents. Your job is to find out and record three characteristics of each person you interview: their age, political ideology, and newspaper reading habits. The natural human tendency would be to record these attributes in words. For example, you might describe a respondent this way: "The respondent is 22 years old,

ideologically moderate, and reads the newspaper about once a week." This would be a good thumbnail description, easily interpreted by another person. To SPSS, though, these words would not make sense.

Whereas people excel at recognizing and manipulating words, SPSS excels at recognizing and manipulating numbers. This is why researchers devise a *coding system*, a set of numeric identifiers for the different values of a variable. For one of the above variables, age, a coding scheme would be straightforward: Simply record the respondent's age in number of years, 22. To record information about political ideology and newspaper reading habits for data analysis, however, a different set of rules is needed. For example, the GSS applies the following coding schemes for political ideology (polviews) and newspaper reading habits (news):

Variable Name (GSS)	Response in Words	Numeric Code
Political ideology (polviews)	Extremely liberal	1
	Liberal	2
	Slightly liberal	3
	Moderate	4
	Slightly conservative	5
	Conservative	6
	Extremely conservative	7
Newspaper reading habits (news)	Every day	1
	A few times a week	2
	Once a week	3
	Less than once a week	4
	Never	5

Thus, the narrative profile "the respondent is 22 years old, is politically moderate, and reads the newspaper about once a week" becomes "22 4 3" to SPSS. SPSS doesn't really care what the numbers stand for. As long as SPSS has numeric data, it will crunch the numbers—telling you the mean age of all respondents or the modal level of newspaper reading. It is important, therefore, to provide SPSS with labels for each code so that the software's analytic work makes sense to the user.

INTERPRETING MEASURES OF CENTRAL TENDENCY AND VARIATION

Finding a variable's central tendency is ordinarily a straightforward exercise. Simply read the computer output and report the numbers. Describing a variable's degree of dispersion or variation, however, often requires informed judgment.[1] Here is a general rule that applies to any variable at any level of measurement: A variable has no dispersion if all the cases—states, countries, people, or whatever—fall into the same value of the variable. A variable has maximum dispersion if the cases are spread evenly across all values of the variable. In other words, the number of cases in one category equals the number of cases in every other category.

Central tendency and variation work together in providing a complete description of any variable. Some variables have an easily identified typical value and show little dispersion. For example, suppose you were to ask a large number of U.S. citizens what sort of political system they believe to be the best: democracy, dictatorship, or anarchy. What would be the modal response, or the economic system preferred by most people? Democracy. Would there be a great deal of dispersion, with large numbers of people choosing the alternatives, dictatorship or anarchy? Probably not.

[1] In this chapter we use the terms *dispersion*, *variation*, and *spread* interchangeably.

In other instances, however, you may find that one value of a variable has a more tenuous grasp on the label *typical*. And the variable may exhibit more dispersion, with the cases spread out more evenly across the variable's other values. For example, suppose a large sample of voting-age adults were asked, in the weeks preceding a presidential election, how interested they are in the campaign: very interested, somewhat interested, or not very interested. Among your own acquaintances you probably know a number of people who fit into each category. So even if one category, such as "somewhat interested," is the median, many people will likely be found at the extremes of "very interested" and "not very interested." In this instance, the amount of dispersion in a variable—its degree of spread—is essential to understanding and describing it.

We can describe the central tendency and dispersion of any variable, but the tools and terminology used to describe a variable depend on the variable's level of measure. The lower the level of measure, the more limited our toolkit for describing central tendency and dispersion. These and other points are best understood by working through some guided examples. In the next section, we'll show you how to use SPSS to describe a nominal-level variable (the lowest level of measurement).

DESCRIBING NOMINAL VARIABLES

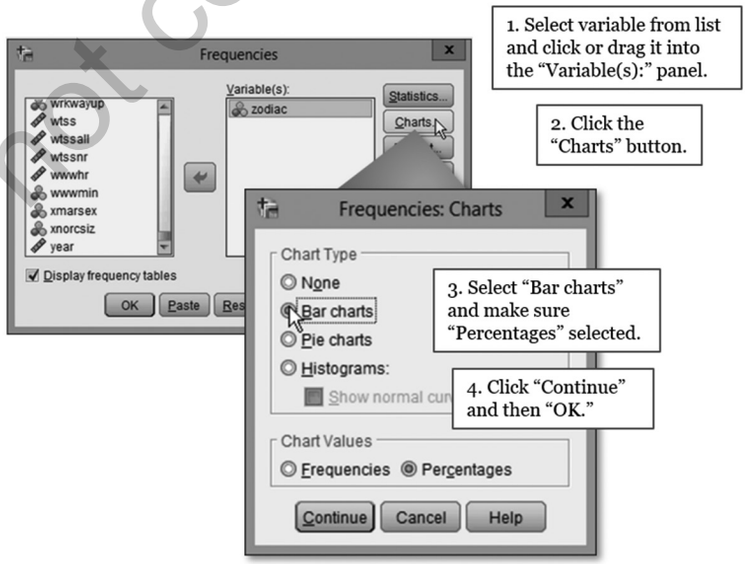
Screencast

Analyze a Variable with Dispersion

For this and the next few analyses, you will use the GSS dataset. Open the GSS dataset by double-clicking the GSS.sav file or, if you already have SPSS running, select File ▶ Open ▶ Data and locate GSS.sav. Before you start analyzing this dataset, select Edit ▶ Options in the Data Editor and then click on the General tab. Just as you did when analyzing a dataset in Chapter 1, make sure that the radio buttons in the Variable Lists area are set for "Display names" and "Alphabetical." (If these options are already set, click Cancel. If they are not set, select them, click Apply, and then click OK. Now you are ready to go.)

First, you will obtain a frequency distribution table and bar chart for a nominal-level variable in the GSS dataset, zodiac, which records respondents' astrological signs. Select Analyze ▶ Descriptive Statistics ▶ Frequencies. Scroll down to the bottom of the left-hand list until you find zodiac. Click zodiac into the Variable(s) panel. To the right of the Variable(s) panel, click the Charts button

FIGURE 2-1 Obtaining Frequencies and a Bar Chart (nominal variable)



(Figure 2-1). The Frequencies: Charts dialog appears. In Chart Type, select "Bar charts." In Chart Values, be sure to select "Percentages." Click Continue, which returns you to the main Frequencies window. Make sure "Display frequency tables" is checked in the Frequencies window. Click OK. SPSS runs the analysis.

SPSS has produced two items of interest in the Viewer: a frequency distribution table of respondents' astrological signs and a bar chart of the same information. (The small "Statistics" table isn't of much interest to us but we include it here so you'll see what's in this book in your Viewer.) First, examine the frequency distribution table.

Statistics

Respondents Astrological Sign

N	Valid	2777
	Missing	90

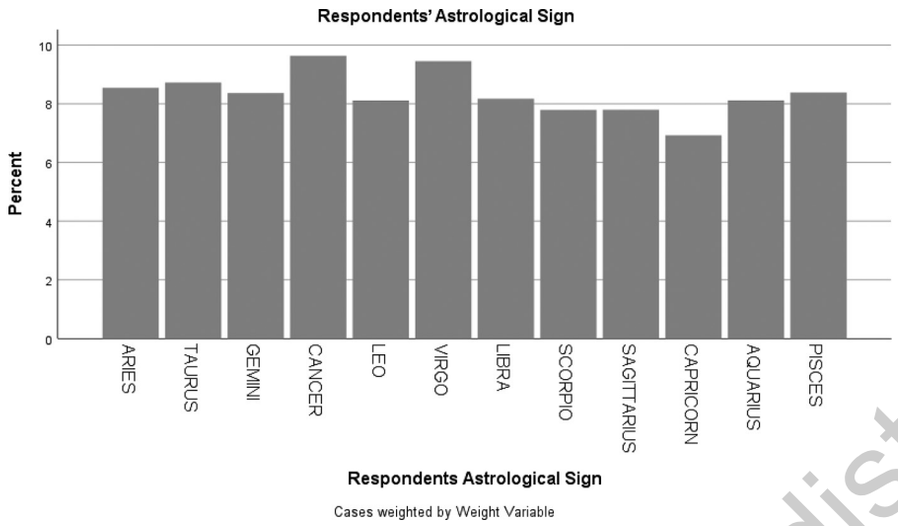
Respondents' Astrological Sign

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	ARIES	237	8.3	8.5	8.5
	TAURUS	242	8.4	8.7	17.3
	GEMINI	232	8.1	8.4	25.6
	CANCER	267	9.3	9.6	35.3
	LEO	225	7.9	8.1	43.4
	VIRGO	262	9.2	9.4	52.8
	LIBRA	227	7.9	8.2	61.0
	SCORPIO	216	7.5	7.8	68.8
	SAGITTARIUS	216	7.5	7.8	76.6
	CAPRICORN	192	6.7	6.9	83.5
	AQUARIUS	225	7.9	8.1	91.6
	PISCES	233	8.1	8.4	100.0
	Total	2777	96.9	100.0	
Missing	System	90	3.1		
Total		2867	100.0		

The value labels for each astrological code appear in the leftmost column, with Aries occupying the top row of numbers and Pisces occupying the bottom row. There are four numeric columns: Frequency, Percent, Valid Percent, and Cumulative Percent. The Frequency column shows raw frequencies, the actual number of respondents having each zodiac sign. Percent is the percentage of *all* respondents, including missing cases, in each category of the variable. Ordinarily the Percent column can be ignored, because we generally are not interested in including missing cases in our description of a variable. Valid Percent is the column to focus on. Valid Percent tells us the percentage of non-missing responses in each value of zodiac. Finally, the Cumulative Percent column reports the percentage of cases that fall in *or below* each value of the variable. For ordinal or interval variables, as you will see, the Cumulative Percent column can provide valuable clues about how a variable is distributed. But for nominal variables like zodiac, which cannot be ranked, the Cumulative Percent column provides no information of value.

Now consider the values in the Valid Percent column more closely. Scroll between the frequency distribution table and the bar chart, which depicts the zodiac variable in graphic form (Figure 2-2).

FIGURE 2-2 Bar Chart of a Nominal-level Variable



What is the mode, the most common astrological sign? For nominal variables, the answer to this question is (almost) always an easy call: Simply find the value with the highest percentage of responses. Virgo is the mode. When it comes to describing the central tendency of nominal-level variables like zodiac, our toolkit is limited to identifying the variable's mode.

Does the zodiac variable have little dispersion or a lot of dispersion? Again, study the Valid Percent column and the bar chart. Apply the following rule: *A variable has no dispersion if the cases are concentrated in one value of the variable; a variable has maximum dispersion if the cases are spread evenly across all values of the variable.* Are most of the cases concentrated in Virgo, or are there many cases in each sign of zodiac? Because respondents are spread out—all astrological signs are about equally represented—you would conclude that zodiac has a high level of dispersion. Looking at the bar chart of zodiac in Figure 2-3, it may be tempting to say the distribution is highest in the middle, but remember that the order of nominal-level values is essentially arbitrary; Virgo is not the middle zodiac sign, so the peak that appears to be in the middle of the bar chart is not a true feature of zodiac's dispersion.

A CLOSER LOOK: WEIGHTING THE GSS AND NES DATASETS

Many of this book's guided examples and exercises use the two survey datasets: the General Social Survey (GSS) and the American National Election Survey (NES). Before proceeding, you need to learn about a feature of these datasets that will require special treatment throughout the book.

In raw form, the GSS and NES datasets are not completely representative of all groups in the population. This lack of representativeness may be intentional (e.g., the American National Election Study purposely oversampled Latino respondents so that researchers could gain insights into the attitudes of this group) or unintentional (e.g., some income groups are more likely to respond to surveys than are other groups). For some SPSS commands, this lack of representativeness does not matter. For most SPSS commands, however, raw survey data produce incorrect results.

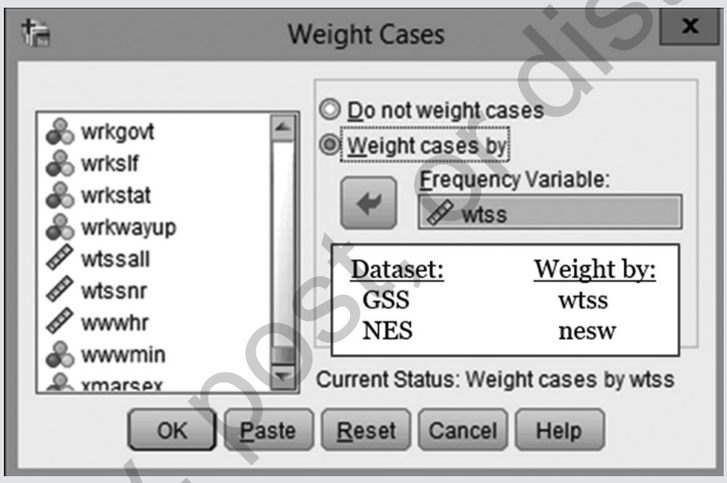
(Continued)

(Continued)

Fortunately, survey designers included the necessary corrective in the NES and GSS dataset: a weight variable. A *weight variable* adjusts for the distorting effect of sampling bias and calculates results that accurately reflect the makeup of the population. If a certain type of respondent is underrepresented in a sample, like young people in a survey conducted by dialing random landline phone numbers, that group's responses are weighted more heavily to make up for being underrepresented. If a certain type of respondent is oversampled, that group's responses are weighted less heavily.

To obtain correct results, *you must specify the weight variable whenever you analyze the GSS or NES datasets.* Otherwise, your analysis will be biased. To weight observations in these datasets to produce nationally representative results, select Data ▶ Weight Cases. When analyzing the GSS, you will specify the weight variable, *wtss* (Figure 2-3). For the NES dataset, the weight variable is *nesw*.

FIGURE 2-3 Weighting Observations in a Dataset



DESCRIBING ORDINAL VARIABLES

In this section, you will analyze and describe two ordinal-level variables in the GSS dataset, one of which has little variation and the other of which is more spread out. These ordinal-level variables examine public opinion on two social policy issues. Opinions on both questions are recorded on 5-point ordinal scales. We will use the same function we did to describe the nominal-level variable zodiac, so click the Analyze tab on the top menu bar of the Viewer and select Analyze ▶ Descriptive Statistics ▶ Frequencies.

SPSS remembers the preceding analysis, so the zodiac variable may still appear in the Variable(s) list. To begin a new descriptive analysis, click zodiac back into the left-hand list.

With your Frequencies dialog window cleared, scroll through the GSS dataset variable list until you find the variable "helppoor" and click on it so it is added to the Variable(s) list. The helppoor variable asks respondents to place themselves on a scale between 1 ("The government should take action to help poor people") and 5 ("People should help themselves"). SPSS should retain your earlier settings for Charts, so accompanying bar charts will appear in the Viewer.[2] Click OK.

[2] If you pressed the Reset button to clear zodiac from the Variable(s) list, you'll need to click the Charts button again to have SPSS produce a bar chart with values specified as percentages.

SPSS produces descriptive statistics for the helppoor variable. To better understand how people responded to the helppoor question, we'll look at the variable's frequency distribution table and bar chart (Figure 2-4).

Statistics

Should Govt Improve Standard of Living?

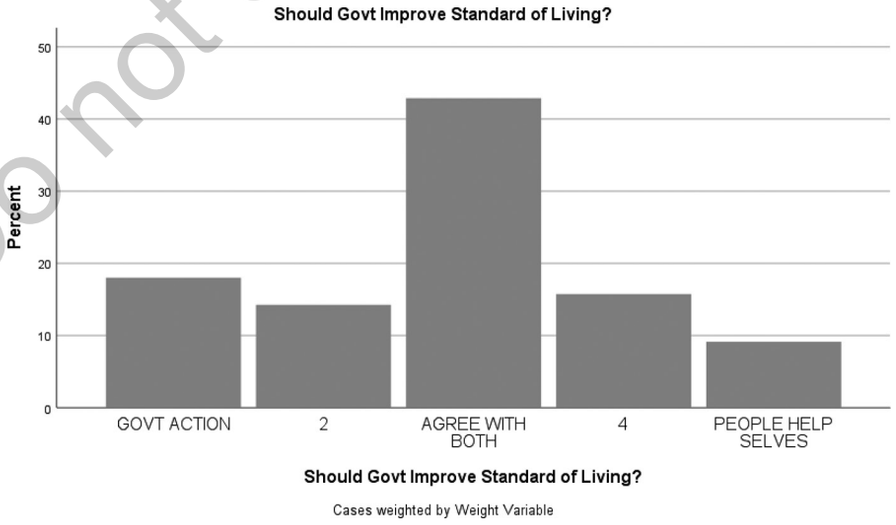
N	Valid	1882
	Missing	985

Should Govt Improve Standard of Living?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	GOVT ACTION	339	11.8	18.0	18.0
	2	268	9.4	14.2	32.3
	AGREE WITH BOTH	807	28.2	42.9	75.1
	4	296	10.3	15.7	90.9
	PEOPLE HELP SELVES	172	6.0	9.1	100.0
	Total	1882	65.7	100.0	
Missing	System	985	34.3		
Total		2867	100.0		

Because helppoor is an ordinal variable, you can use both its mode and its median to describe central tendency. Its mode, clearly enough, is the response "Agree with both," which contains 42.9 percent of the cases. (If you get a different percentage in this category, make sure you've weighted observations properly.) What about the median? This is where the Cumulative Percent column of the frequency distribution comes into play. *The median for any ordinal (or interval) variable is the*

FIGURE 2-4 Bar Chart of an Ordinal Variable with Low Dispersion



category below which 50 percent of the cases lie. Is the first category, "Govt action," the median? No, this category contains fewer than half of the cases. How about the next higher category? No, again. The Cumulative Percent column still has not reached 50 percent. The median occurs in the "Agree with both" category (cumulative percentage, 75.1).

Now consider the question of whether helppoor has a high degree of dispersion or a low degree of dispersion. If helppoor had a high level of variation, then the percentages of respondents in each response category would be roughly equal, much like the zodiac variable that you analyzed earlier. So, roughly one-fifth of the cases would fall into each of the five response categories: 20 percent in "Gov't action," 20 percent in response category "2," 20 percent in "Agree with both," 20 percent in response category "4," and 20 percent in "People help selves." If helppoor had no dispersion, then all the cases would fall into one value. That is, one value would have 100 percent of the cases, and each of the other categories would have 0 percent. Which of these two scenarios comes closest to describing the actual distribution of respondents across the values of helppoor: the equal-percentages-in-each-category, high variation scenario, or the 100-percent-in-one-category, low variation scenario? It seems clear that helppoor is a variable with a relatively low degree of dispersion. "Agree with both," with 42.9 percent of the responses, contains nearly three times as many cases as its nearest rival ("Gov't action") and more than three times as many cases as any of the other response categories.

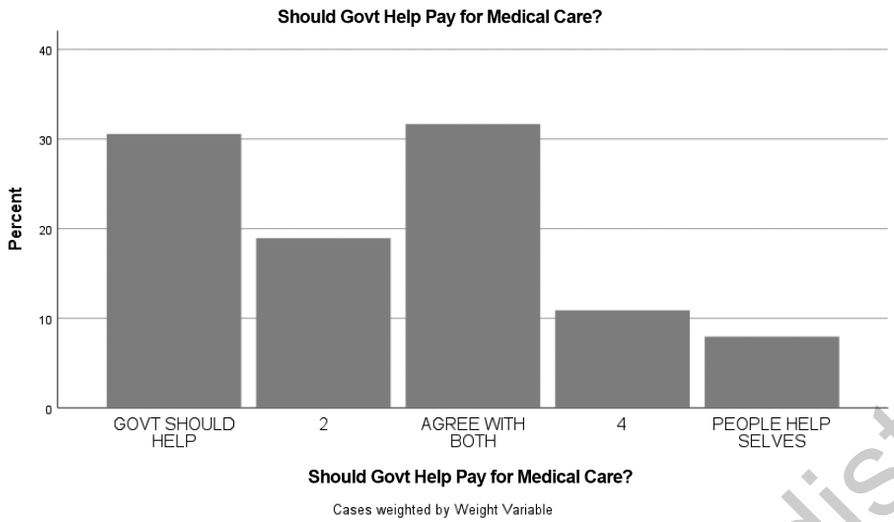
Now contrast helppoor's distribution with the distribution of the helpsick variable. The "helpsick" variable, using a similar 5-point scale, asks respondents about government responsibility or individual responsibility for medical care. You can produce a frequency distribution table and bar chart for helpsick (Figure 2-5) the same way we did to generate descriptive statistics for the helppoor variable. Review the preceding paragraphs as necessary to do this analysis.

Should Govt Help Pay for Medical Care?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	GOVT SHOULD HELP	579	20.2	30.6	30.6
	2	358	12.5	18.9	49.5
	AGREE WITH BOTH	599	20.9	31.7	81.2
	4	206	7.2	10.9	92.1
	PEOPLE HELP SELVES	150	5.2	7.9	100.0
	Total	1893	66.0	100.0	
Missing	System	974	34.0		
Total		2867	100.0		

Interestingly, helpsick has the same mode as helppoor ("Agree with both," with 31.7 percent of the cases), and the same median (again, "Agree with both," where the cumulative percentage exceeds 50.0). Yet with helppoor it seemed reasonable to say that "Agree with both" was the typical response. Would it be reasonable to say that "Agree with both" is helppoor's typical response? No, it would not. Notice that, unlike helppoor, respondents' values on helpsick are more spread out, with sizable numbers of responses falling in the first value ("Gov't action," with 30.6 percent), making it a close rival to "Agree with both" for the distinction of being the modal opinion on this issue. Clearly, the public is more divided—more widely dispersed—on the question of medical assistance than on the question of assistance to the poor.

FIGURE 2-5 Bar Chart of an Ordinal Variable with High Dispersion



A CLOSER LOOK: ANALYZING TWO VARIABLES AT ONCE

In the preceding section, we demonstrated how to describe two ordinal-level variables, helppoor and helpsick, using frequency distribution tables and bar charts. We analyzed one variable at a time to keep things simple as you learn a new skill and to draw your attention to the differences between these two variables. In the future, if you are describing the distributions of two or more variables, you may prefer to analyze multiple variables at once to work more efficiently. Watch the "Analyze Two Variables" screencast to learn how to analyze two (or more) variables at once.

Watch Screencast
Analyze Two Variables

USING THE CHART EDITOR TO MODIFY GRAPHICS

SPSS permits the user to modify the content and appearance of any graphic object it produces in the Viewer using the Chart Editor. The user invokes the Editor by double-clicking on graphical output in the Viewer, edits the graphic using the Chart Editor tool, and closes the Chart Editor to return to the Viewer. Changes made to a graphic in the Chart Editor are recorded automatically in the Viewer.

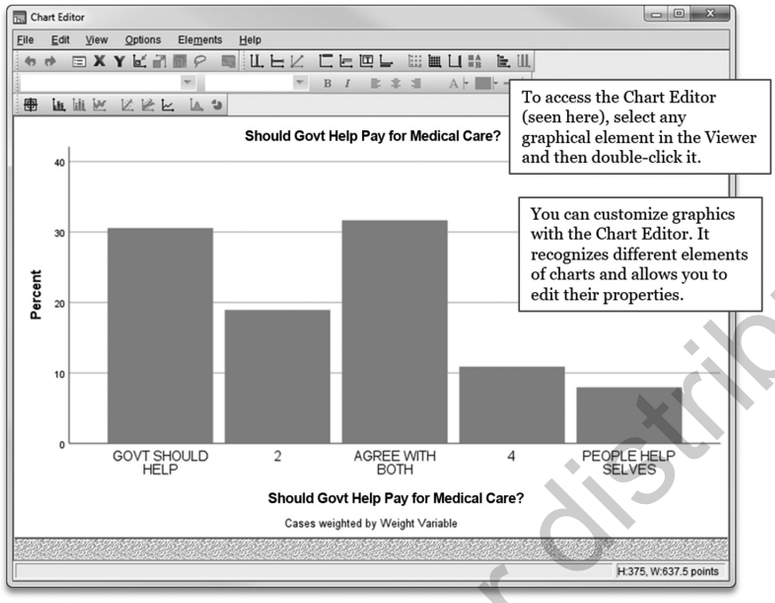
In this section, we show how you can use the Chart Editor to edit the bar chart you just created to show the distribution of public opinion on paying for medical care for sick people. We'll show how to change labels on the chart and also change the style and color of the bars. (The default style and color is rather uninspired, and it doesn't print well.)

In the Viewer, place the cursor anywhere on the bar chart and double-click. SPSS opens the Chart Editor (Figure 2-6).

The Chart Editor recognizes the elements that make up the bar chart. It recognizes some elements as text. These elements include the axis titles and the value labels for the categories of helpsick. It recognizes other elements as graphic, such as the bars in the bar chart. First, we'll edit a text element, the title on the vertical axis. Then we will modify a graphic element, the color of the bars.

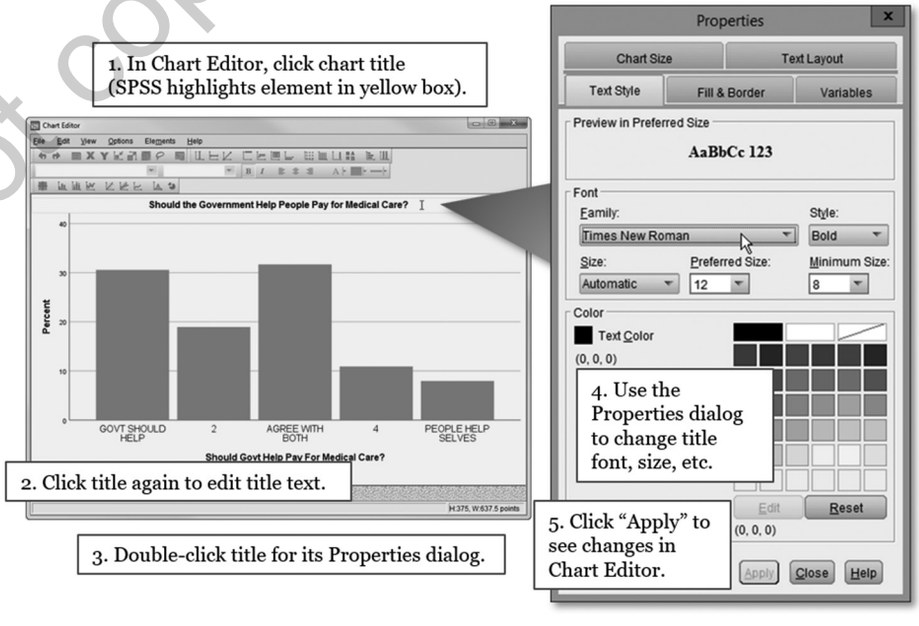
Place the cursor anywhere on the main title "Should Govt Help Pay For Medical Care?" and single-click it. SPSS selects the chart title. With the cursor still placed on the title, single-click again. SPSS moves the text into editing mode inside the chart (see the left side of Figure 2-7). The default

FIGURE 2-6 The Chart Editor



title is fine for understanding the distribution of this variable, but we could improve it to present the information to others. Edit the title so it reads "Should the Government Help People Pay for Medical Care?" We want the bar chart to communicate its information as clearly as possible. While you're using the Chart Editor to edit chart text, you can delete the redundant *x*-axis label and provide more descriptive *x*-axis labels in place of the numbers "2" and "4."

FIGURE 2-7 Editing the Bar Chart Title



You can also use the Text Style menu in the Properties window that pops up automatically when you double-click a graphic element in the Chart Editor (see the right side of Figure 2-7) to change the font style of the main title and *x*-axis labels from the plain, default sans serif font to something more stylish. If you're going to use an SPSS graphic in a paper or presentation, you may want to match the font used in the graphic with the font used in the paper or presentation. When you change the font of a text element, change the font of all the other text elements to match so your graphic doesn't become a hodgepodge of text styles. Apply your changes to the chart text.

Now click on one of the vertical bars. The editor selects all the bars. As before, you can double-click an element in the Chart Editor to summon the associated Properties dialog window. Alternatively, you can select the element and press the "Show Properties Window" button located near the upper-left corner of the Chart Editor window as we show in Figure 2-8. This opens the Properties window, the most powerful editing tool in the Chart Editor's arsenal. (*Special note:* If you plan to do a lot of editing, it is a good idea to open the Properties window soon after you enter the Chart Editor. Each time you select a different text or graphic element with the mouse, the Properties window changes, displaying the editable properties of the selected element.)

The options for editing graphical elements like the bars in a bar chart are plentiful. You can change their color, adjust their order, and make them bigger or smaller. The "Depth & Angle" tab of the bar properties provides an option that dramatically transforms the humble bar chart into a visually interesting graphic: a 3-D effect. Select the 3-D option (see Figure 2-8) and apply it to the bar chart. You'll see the difference this option makes in the Chart Editor. If you close the Chart Editor, the finished product appears in the Viewer (Figure 2-9).

DESCRIBING INTERVAL VARIABLES

Let's now turn to the descriptive analysis of interval-level variables. An interval-level variable represents the most precise level of measurement. Unlike nominal variables, whose values stand for categories, and ordinal variables, whose values can be ranked, the values of an interval variable *tell you the exact quantity of the characteristic being measured*. For example, age qualifies as an interval-level variable because its values impart each respondent's age in years.

Screencast

Frequency Analysis with an Interval Variable

FIGURE 2-8 Using the Properties Window to Edit Bars in a Bar Chart

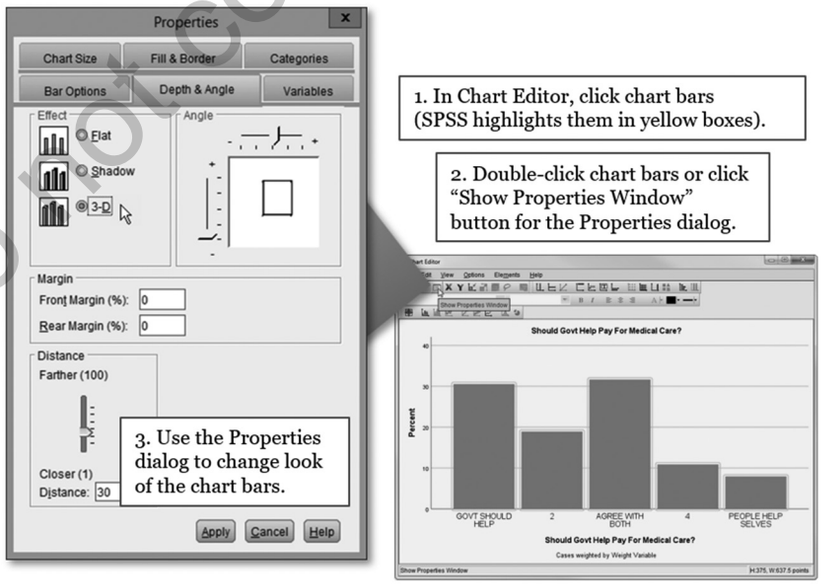
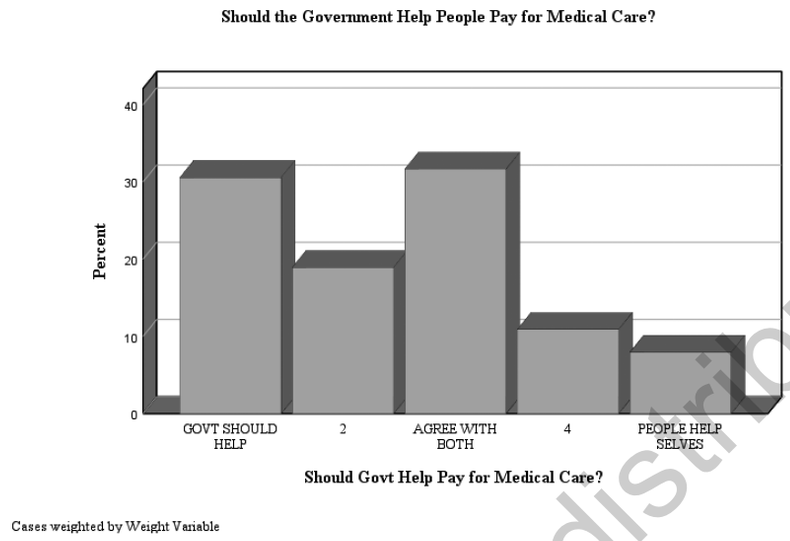


FIGURE 2-9 Edited Bar Chart in the Viewer



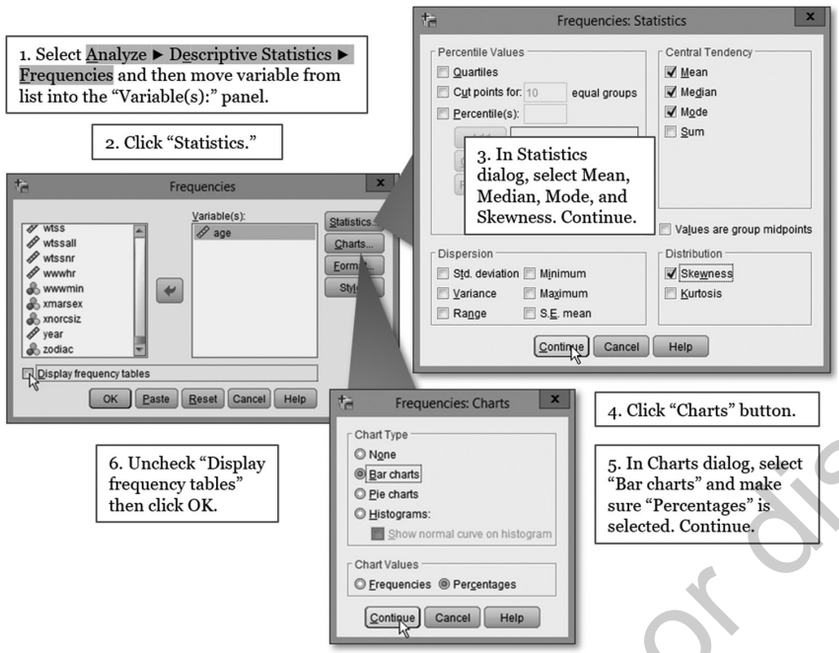
Because interval variables have the most precision, they can be described more completely than can nominal or ordinal variables; we have a relatively large toolkit available for describing variables measured at the interval level. For any interval-level variable, you can report its mode, median, and arithmetic average, or *mean*. In addition to these measures of central tendency, you can make more sophisticated judgments about variation. Specifically, you can determine if an interval-level distribution is *skewed*.

Skewness refers to the symmetry of a distribution. If a distribution is not skewed, the cases tend to cluster symmetrically around the mean of the distribution, and they taper off evenly for values above and below the mean. If a distribution is skewed, by contrast, one tail of the distribution is longer and skinnier than the other tail. Distributions in which some cases occupy the higher values of an interval variable—distributions with a skinnier right-hand tail—have a *positive skew*. By the same token, if the distribution has some cases at the extreme lower end—the distribution has a skinnier left-hand tail—then the distribution has a *negative skew*. Skewness affects the mean of the distribution. A positive skew tends to "pull" the mean upward; a negative skew pulls it downward. However, skewness has less effect on the median. Because the median reports the middlemost value of a distribution, it is not tugged upward or downward by extreme values. *For badly skewed distributions, it is a good practice to use the median instead of the mean in describing central tendency.*

A step-by-step analysis of a GSS variable, age, will clarify these points. Select Analyze ▶ Descriptive Statistics ▶ Frequencies. If helppoor and helpsick are still in the Variable(s) list, click them back into the left-hand list. Click age into the Variable(s) list. You may notice that the icon next to the age variable looks different from the icon next to zodiac, helppoor, and helpsick; the icons signify the variable's level of measurement.

So far, this procedure is the same as in your analysis of zodiac, helppoor, and helpsick. When you are running a frequencies analysis of an interval-level variable, however, you need to adjust the settings for the Frequency analysis to get proper results. Here's a must-do: Click the Statistics button in the Frequencies window, as shown in Figure 2-10. The Frequencies: Statistics window appears. In the Central Tendency panel, click the boxes next to Mean, Median, and Mode. In the Distribution panel, click Skewness. Click Continue, returning to the main Frequencies window. Now click the Charts button. In the Charts dialog, make sure that "Bar charts" (under Chart Type) and "Percentages" (under Chart Values) are selected. Click Continue.

FIGURE 2-10 Requesting Statistics for an Interval Variable



While you're in the main Frequencies window, here's something you may want to do before you execute the analysis: *Uncheck* the box next to "Display frequency tables," appearing at the foot of the left-hand list.[3] For interval-level variables, like age, that have many categories, a frequency distribution table can run several output pages and is not very informative. Unchecking the "Display frequency tables" box suppresses the frequency distribution. Click OK.

SPSS analyzes the age variable and outputs the requested statistics and bar chart (Figure 2-11) into the Viewer. Most of the entries in the Statistics table are familiar to you: valid number of cases; number of missing cases; and mean, median, and mode. In addition, SPSS reports values for skewness and a statistic called standard error of skewness. When a distribution is perfectly symmetrical—no skew—it has skewness equal to 0. If the distribution has a skinnier right-hand tail—positive skew—then skewness will be a positive number. A skinnier left-hand tail, logically enough, returns a negative number for skewness.

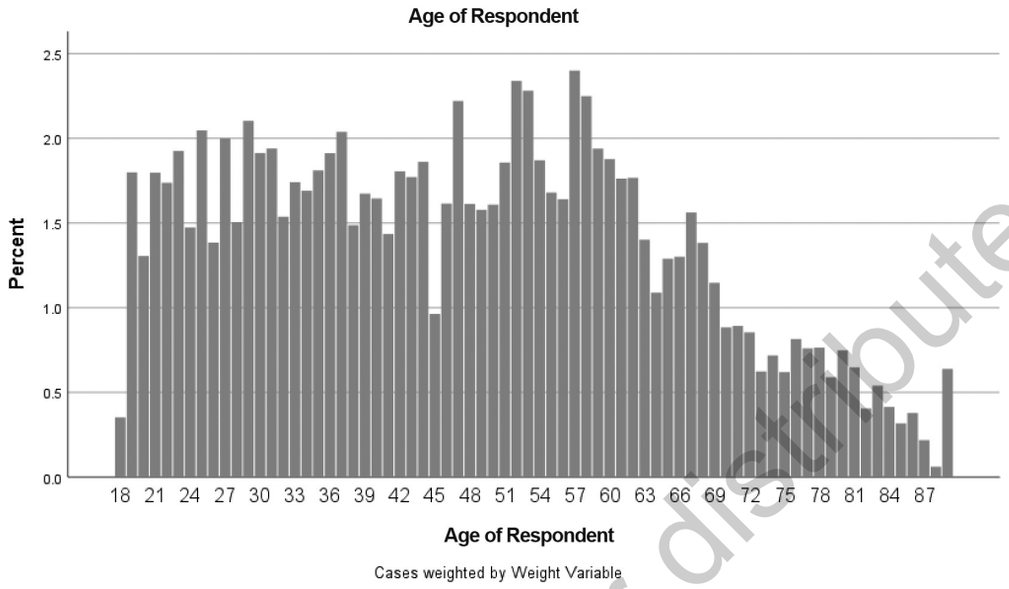
Statistics

Age of Respondent

N	Valid	2855
	Missing	12
Mean		47.56
Median		47.00
Mode		57
Skewness		.233
Std. Error of Skewness		.046

[3] A general guide: If the interval-level variable you are analyzing has 15 or fewer categories, go ahead and obtain the frequency distribution. If it has more than 15 categories, suppress the frequency distribution.

FIGURE 2-11 Bar Chart of the Interval-level Age Variable



For the age variable, the skewness statistic is positive (.233). This suggests that the distribution has a skinnier right-hand tail—a feature that is confirmed by the shape of the bar chart. Note also that the mean (47.56 years) is slightly higher than the median (47 years), a situation that often—although not always—indicates a positive skew.[4] Even so, the mean and median are less than 1 year apart. You have to exercise judgment, but in this case, it would not be a distortion of reality to use the mean instead of the median to describe the central tendency of the distribution.[5]

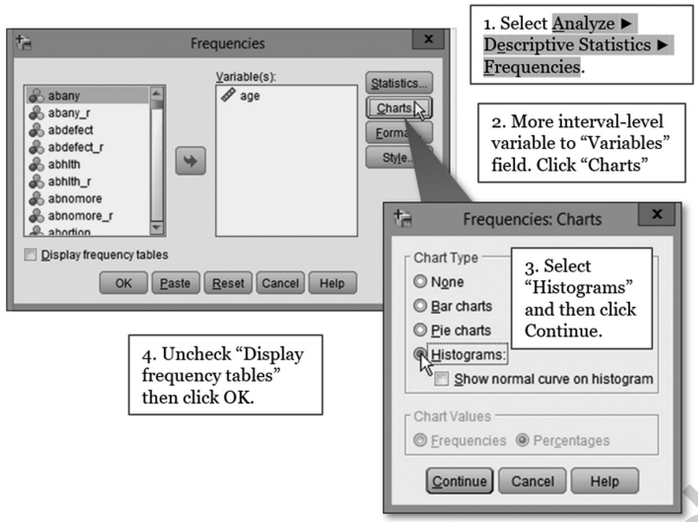
All the guided examples thus far have used bar charts for graphic support. For nominal and ordinal variables, a bar chart should always be your choice. For interval variables, however, you may want to ask SPSS to produce a histogram instead. What is the difference between a bar chart and a histogram? A bar chart displays each value of a variable and shows you the percentage (alternatively, the raw number) of cases that fall into each category. A histogram is similar to a bar chart, but instead of displaying each discrete value, it collapses categories into ranges (called bins), resulting in a compact display. Histograms are sometimes more readable and elegant than bar charts. Most of the time a histogram will work just as well as a bar chart in summarizing an interval-level variable. For interval variables with many unique values, a histogram is the graphic of choice. (Remember: For nominal or ordinal variables, you always want a bar chart.)

So that you can become familiar with histograms, run the analysis of age once again—only this time ask SPSS to produce a histogram instead of a bar chart. Select Analyze ▶ Descriptive Statistics ▶ Frequencies. Make sure age is still in the Variable(s) list. Click Statistics, and then *uncheck* all the boxes: Mean, Median, Mode, and Skewness. Click Continue. Click Charts, and then select the Histograms radio button in Chart Type. Click Continue. For this analysis, we do not

[4] Paul T. von Hippel, "Mean, Median, and Skew: Correcting a Textbook Rule," *Journal of Statistics Education* 13, no. 2 (2005). "Many textbooks teach a rule of thumb stating that the mean is right of the median under right skew, and left of the median under left skew. This rule fails with surprising frequency." See http://www.amstat.org/publications/jse/v13n2/vonhippel.html.

[5] For demographic variables that are skewed, median values rather than means are often used to give a clearer picture of central tendency. One hears or reads reports, for example, of median family income or the median price of homes in an area.

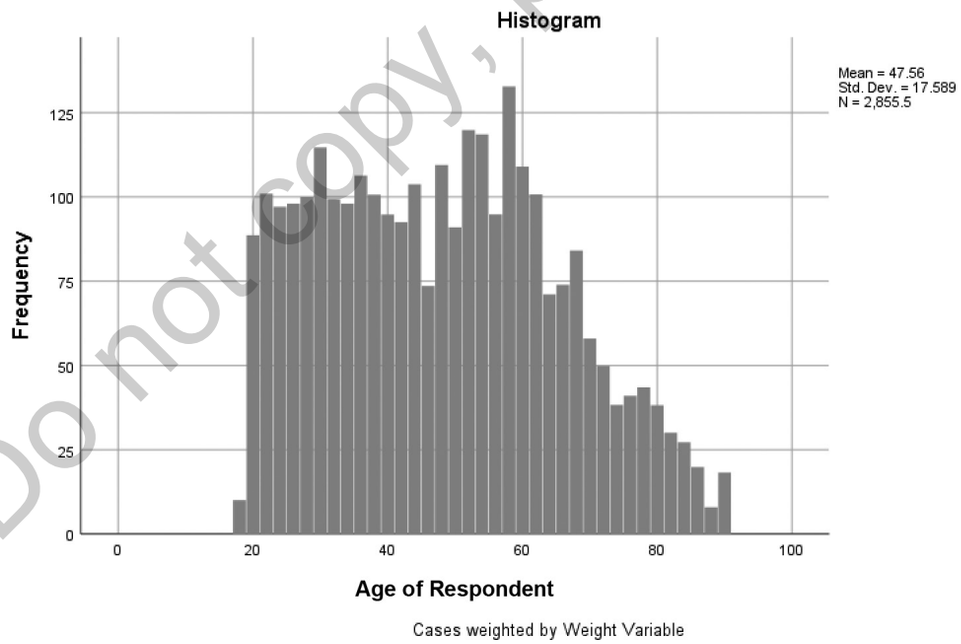
FIGURE 2-12 Creating a Histogram of an Interval Variable



need a frequency distribution table. In the Frequencies window, uncheck the "Display frequency tables" box. (Refer to Figure 2-12.) Click OK.[6]

This is a bare-bones run. SPSS reports its obligatory count of valid and missing cases, plus a histogram for age (Figure 2-13). On the histogram's horizontal axis, notice the tick marks, which are spaced at 20-year intervals. SPSS has compressed the data so that each bar represents about 2

FIGURE 2-13 Histogram of the Interval-level Age Variable



[6] Alternatively, you can select Graphs ▶ Legacy Dialogs ▶ Histogram to produce a histogram. Move age into the Variable field and click OK. This procedure should produce the same output as the method described in the text.

years of age rather than 1 year of age. Now scroll up the Viewer to the bar chart of age, which you produced in the preceding analysis. Notice that the histogram has smoothed out the choppiness of the bar chart, though it still captures the essential qualities of the age variable.

Statistics

Age of Respondent

N	Valid	2855
	Missing	12

OBTAINING CASE-LEVEL INFORMATION WITH CASE SUMMARIES

Screencast

Analyze Case Summaries

When you analyze a large survey dataset, as you have just done, you generally are not interested in how respondent X or respondent Y answered a particular question; they're just some random people who happened to participate in the survey. Rather, you want to know how the entire sample of respondents distributed themselves across the response categories of a variable (for this purpose, their randomness is vitally important). Sometimes, however, you gather data on particular cases because the cases are themselves inherently important.

When you work with the States dataset (states.sav) and the World dataset (world.sav), you may want to describe cases beyond the relative anonymity of Frequencies analysis and find out where particular states or countries "are" on an interesting variable. To obtain case-level information, select Analyze ▶ Reports ▶ Case Summaries. This SPSS procedure is readymade for such elemental insights.

Suppose you are interested in identifying states that have the most/fewest laws restricting access to abortions. To begin this guided example, close the GSS dataset and open the States dataset. The States dataset contains a variable named abortlaw17. This variable records the number of legal restrictions on abortion access in each state in 2017 (out of 14 possible restrictions). Exactly which states impose the most restrictions? Which states impose the fewest restrictions? Where does your state fall on the list? Case Summaries can quickly answer questions like these. SPSS will sort states based on a "grouping variable" (in this example, abortlaw17) and then produce a report telling you which states are in each group.

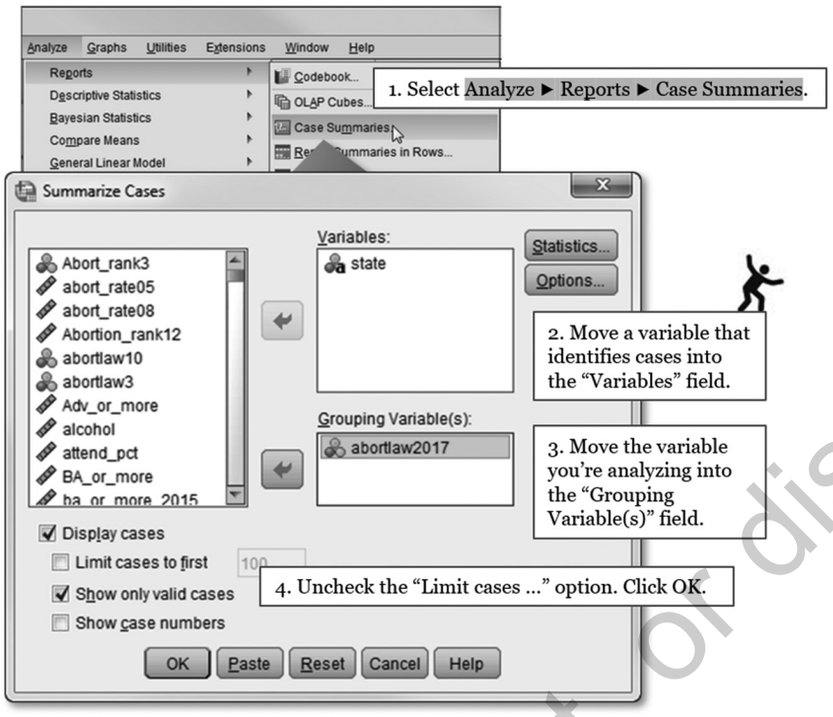
With the States dataset open, click Analyze ▶ Reports ▶ Case Summaries.

To conduct the desired analysis, you need to do three things in the Summarize Cases window (see Figure 2-14):

1. Click the variable containing the cases' identities into the Variables window. In the States dataset, this variable is named state, which is simply the name of each state.
2. Click the variable you are interested in analyzing, abortlaw17, into the Grouping Variable(s) window.
3. Uncheck the "Limit cases to first. . ." option. This won't affect the analysis of state abortion laws because there are fewer than 100 states, but if this box is left checked when you analyze the World dataset, SPSS will limit the analysis to the first 100 countries and produce an incomplete analysis.

Click OK and consider the output. SPSS sorts the cases on the grouping variable, abortlaw17, and tells us which state is associated with each value of abortlaw17. For example, Vermont, with 0 legal restrictions on abortion access, is the state with the fewest restrictions on access to abortions. Which states impose the most restrictions? Scroll to the bottom of the tabular output. With 13 restrictions, Kansas and Oklahoma are tied for imposing the most restrictions.

FIGURE 2-14 Obtaining Case Summaries



Case Processing Summary

	Cases					
	Included		Excluded		Total	
	N	Percent	N	Percent	N	Percent
State Name * Number of restrictions on abortion	50	100.0%	0	0.0%	50	100.0%

Case Summaries

			State Name	
Number of restrictions on abortion	0	1	Vermont	
		Total N		1
	2	1	Colorado	
		2	New Hampshire	
		3	Oregon	
		Total N		3
	3	1	California	
		2	Connecticut	
		3	New York	
		4	Washington	
		Total N		4
	4	1	Hawaii	
		2	New Jersey	
		3	New Mexico	
		4	Rhode Island	
		5	West Virginia	
		Total N		5
	5	1	Alaska	
		2	Illinois	
		3	Maine	
		4	Maryland	
		5	Massachusetts	
		6	Montana	
		Total N		6
	6	1	Delaware	
		2	Nevada	
		3	Wyoming	
		Total N		3
	7	1	Alabama	
				[Output omitted]
		4	Wisconsin	
		Total N		4
	9	1	Florida	
		2	Georgia	
		3	Kentucky	
		4	Mississippi	
		5	Nebraska	
		6	North Dakota	
		7	Pennsylvania	
		Total N		7
	10	1	Arizona	
		2	Arkansas	
		3	Michigan	
		4	Ohio	
		5	South Carolina	
		6	South Dakota	
		7	Tennessee	
		8	Texas	
		9	Utah	
		10	Virginia	
		Total N		10
	11	1	Louisiana	
		2	Missouri	
		Total N		2
	12	1	Indiana	
		Total N		1
	13	1	Kansas	
		2	Oklahoma	
		Total N		2
	Total	N		50

Note: The state names have been edited for spacing.

Name: _____ Date: _____

E-mail: _____ Section: _____

CHAPTER 2 EXERCISES

1. (Data set: GSS. Variables: science_quiz, wtss.) The late Carl Sagan once lamented, "We live in a society exquisitely dependent on science and technology, in which hardly anyone knows anything about science and technology." This is a rather pessimistic assessment of the scientific acumen of ordinary Americans. Sagan seemed to be suggesting that the average level of scientific knowledge is quite low and that most people would fail even the simplest test of scientific facts.

   The GSS dataset contains science_quiz, which was created from ten true-false questions testing respondents' knowledge of basic scientific facts. Values on science_quiz range from 0 (the respondent did not answer any of the questions correctly) to 10 (the respondent correctly answered all ten).[7]

   A. Obtain a frequency distribution table of science_quiz. Fill in the table that follows. Be sure to weight observations with the wtss variable.

science_ quiz Score	Frequency*	Percent	Cumulative Percent
0	2	?	?
1	9	?	?
2	13	?	?
3	38	?	?
4	54	?	?
5	66	?	?
6	80	?	?
7	78	?	?
8	60	?	?
9	45	?	?
10	21	?	100.00%
Total	465	100.00%	

*Weighted frequencies.

   B. When you use the Analyze ▶ Descriptive Statistics ▶ Frequencies procedure, click the Statistics button and ask SPSS to report the mean, median, and skewness of the science_quiz variable. The science_quiz variable has a mean equal to _____, a median equal to _____, and a skewness equal to _____.

   C. Create a bar chart for science_quiz by clicking the Charts button and requesting a bar chart. Print the bar chart. (Alternatively, you can create a histogram, but there is no need to group observations into binned values of science_quiz values.)

   D. Exercise your judgment. What would be the more accurate measure of science_quiz's central tendency: the mean or the median? (circle one)

      mean median

   E. Briefly explain your choice in D.

      _____

      _____

      _____

      _____

   F. According to conventional academic standards, any science_quiz score of 5 or lower would be an F, a failing grade. A score of 6 would be a grade of D, a 7 would be a C, an 8 a B, and scores of 9 or 10 would be an A. Based on these standards, about what percentage of people got passing grades on science_quiz? (circle one)

      About 30 percent About 40 percent
      About 50 percent About 60 percent

      What percentage got an A on science_quiz? (circle one)

      About 5 percent About 10 percent
      About 15 percent About 20 percent

[7] The science_quiz variable was created by summing the number of correct responses to the following questions (all are in true-false format, except for earthsun): The center of the Earth is very hot (General Social Survey variable, hotcore); it is the father's gene that decides whether the baby is a boy or a girl (boyorgrl); electrons are smaller than atoms (electron); the universe began with a huge explosion (bigbang); the continents on which we live have been moving their locations for millions of years and will continue to move in the future (condrift); human beings, as we know them today, developed from earlier species of animals (evolved); does the Earth go around the sun, or does the sun go around the Earth (earthsun); all radioactivity is manmade (radioact); lasers work by focusing sound waves (lasers); and antibiotics kill viruses as well as bacteria (viruses).

2. (Dataset: World. Variables: women13, country.) What percentage of members of the U.S. House of Representatives are women? In 2013 the number was 17.8 percent, according to the Inter-Parliamentary Union, an international organization of parliaments.[8] How does the United States compare to other democratic countries? Is 17.8 percent comparatively low, comparatively high, or average for a typical national legislature?

   A. The World dataset contains women13, the percentage of women in the lower house of the legislature in each of ninety democracies. Obtain summary statistics for the women13 variable. Fill in the table that follows.

Statistics for women13 Variable	
Mean	?
Median	?
Skewness	?

   B. Examine the results of the summary analysis. Recall that 17.8 percent of U.S. House members are women. Now, consider the following statement: "The percentage of women in the U.S. House is about average for a democratic country." Is this statement accurate? Answer yes or no, and explain your reasoning.

_____

_____

_____

_____

   C. Suppose a women's advocacy organization vows to support female congressional candidates so that the U.S. House might someday "be ranked among the top 10 percent of democracies in the percentage of female members." According to the percentiles column of the summary analysis, to meet this goal women would need to constitute about what percentage of the House? (circle one)

About 25 percent About 40 percent
About 50 percent

   D. Create a histogram of women13. Print the histogram.

   E. Run Analyze ▶ Reports ▶ Case Summaries. Click Country into the Variables box and women13 into the Grouping Variable(s) box. Make sure to uncheck the box next to "Limit cases to first 100." Examine the output.

The five countries with the *lowest percentages* of women legislators are

1. _____
2. _____
3. _____
4. _____
5. _____

The five countries with the *highest percentages* of women legislators are

1. _____
2. _____
3. _____
4. _____
5. _____

3. (Dataset: GSS. Variables: femrole, wtss.) Two pundits are arguing about the general public's views on the role of women in the home and in politics.

Pundit 1: "Our society has a minority of traditionally minded individuals who think that the proper 'place' for women is taking care of the home and caring for children. This small but vocal group of traditionalists aside, the typical adult supports the idea that women belong in work and in politics."

Pundit 2: "Poppycock! It's just the opposite. The extremist feminist crowd has distorted the overall picture. The typical view among most citizens is that women should be in the home, not in work and politics."

   A. Dataset GSS (file name: gss.sav) contains femrole, an interval-level variable that measures respondents' attitudes toward women in society and politics. Scores can range from 0 (women belong in the home) to 9 (women belong in work and politics).

If Pundit 1 is correct, femrole will have (circle one)

a negative skew. no skew. a positive skew.

If Pundit 2 is correct, femrole will have (circle one)

a negative skew. no skew. a positive skew.

If Pundit 1 is correct, femrole's mean will be (circle one)

lower than its median. the same as its median.
higher than its median.

If Pundit 2 is correct, femrole's mean will be (circle one)

lower than its median. the same as its median.
higher than its median.

   B. Perform a frequencies analysis of femrole. Obtain the mean, median, and mode, as well as skewness. Obtain a bar chart with percentages. Fill in the table that follows.

[8] See the Inter-Parliamentary Union website (https://www.ipu.org).

Statistics for femrole Variable	
Mean	?
Median	?
Mode	?
Skewness	?

C. Create a bar chart of femrole. Be sure to use sample weights so the distribution is nationally representative. Override the default bar fill color with a color of your choice. Print the bar chart.

D. Consider the evidence you obtained in parts B and C. Based on your analysis, whose assessment is more accurate? (circle one)

Pundit 1's Pundit 2's

Citing *specific evidence* obtained in parts B and C, explain your reasoning.

_____

_____

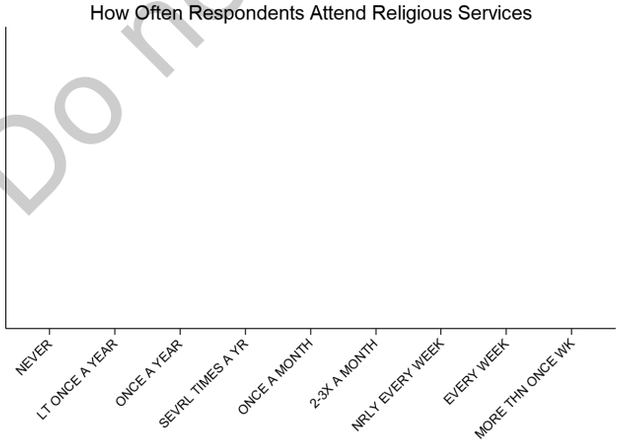
_____

_____

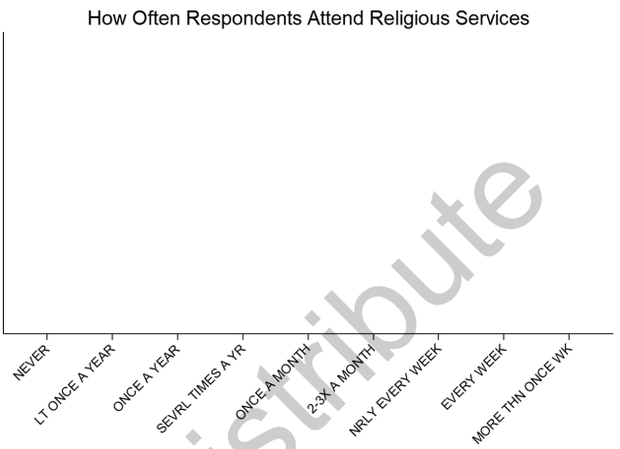
_____

4. (Dataset: GSS. Variables: attend, wtss.) The GSS dataset (file name: gss.sav) provides a rich array of variables that permit scholars to study religiosity in the adult population. The GSS dataset contains attend, a 9-point ordinal scale that measures how often respondents attend religious services. Values can range from 1 ("Never") to 9 ("More than once a week").

A. The shell of a bar chart is given below. The categories of attend appear along the horizontal axis. What would a bar chart of attend look like if this variable had maximum dispersion? Sketch inside the axes a bar chart that would depict maximum dispersion.



B. What would a bar chart of attend look like if this variable had no dispersion? Sketch inside the axes a bar chart that would depict no dispersion.



C. Obtain frequencies output and a bar chart for attend. In the main Frequencies window, make sure that the "Display frequencies table" box is checked. In Statistics, see that all the boxes are unchecked. In Charts, request a bar chart with percentages. Based on your results, complete the following table.

attend Value	Frequency*	Percent	Cumulative Percent
Never	711.33	?	?
Less than once a year	167.88	?	?
Once a year	377.98	?	?
Several times a year	316.54	?	?
Once a month	198.52	?	?
2 to 3 times a month	249.13	?	?
Nearly every week	126.56	?	?
Every week	498.02	?	?
More than once a week	204.03	?	100.00%
Total	2,850	100.00%	

*Weighted frequencies rounded to two decimal places

D. Print the bar chart you obtained for the last part of this exercise.

E. Based on your examination of the frequency distribution,

the mode of attend is _____.

the median of attend is _____.

F. Based on your examination of the frequency distribution and bar chart, you would conclude that attend has (circle one)

low dispersion. high dispersion.

5. (Dataset: NES. Variables: immig_chldrn, grass, pres_job, nesw.) We frequently describe public opinion by referring to how citizens distribute themselves on a political issue. *Consensus* is a situation in which a large majority, 60–70 percent of the public, holds the same position, or very similar positions, on an issue. *Dissensus* is a situation in which opinion is spread out evenly across all positions on an issue. *Polarization* refers to a configuration of opinion in which people are split between two extreme poles of an issue, with only a few individuals populating the more moderate, middle-of-the-road positions.

In this exercise you will decide whether consensus, dissensus, or polarization best describes public opinion, as measured by three NES variables: opinions about allowing people brought to this country illegally as children to stay in the United States (immig_chldrn), opinions on whether marijuana should be legalized (grass), and opinions about how well the president is performing his job (pres_job). The question regarding children brought illegally is measured on a 6-point scale, from "Should send back—favor a great deal" (point 1) to "Should allow to stay—favor a great deal" (point 6). The marijuana legalization question uses a 3-point scale: "Favor" (point 1), "Neither favor nor oppose" (point 2), and "Oppose" (point 3). Presidential job approval is measured on a 4-point scale, from "Approve strongly" (point 1) to "Disapprove strongly" (point 4).[9]

A. Obtain frequency distributions and bar charts for immig_chldrn, grass, and pres_job. Remember to weight the analyses using nesw. In the tables that follow, write the appropriate percentage next to each question mark.

Send back children brought to U.S. illegally?	Percentage
1. Should send back—favor a great deal	?
2. Should send back—favor a moderate amount	?
3. Should send back—favor a little	?
4. Should allow to stay—favor a little	?
5. Should allow to stay—favor a moderate amount	?
6. Should allow to stay—favor a great deal	?
Total	100.00%

Should marijuana be legal?	Percentage
Favor	?
Neither favor nor oppose	?
Oppose	?
Total	100.00%

Presidential Approval Scale	Percentage
Approve strongly	?
Approve	?
Disapprove	?
Disapprove strongly	?
Total	100.00%

B. Examine the percentages you entered in the tables above. Of the three issues, which one *most closely approximates* consensus? (check one)

- ❑ Sending back children brought to U.S. illegally
- ❑ Legalization of marijuana
- ❑ Presidential approval

Briefly explain your reasoning.

C. Of the three issues, which one *most closely approximates* dissensus? (circle one)

- ❒ Sending back children brought to U.S. illegally
- ❒ Legalization of marijuana
- ❒ Presidential approval

Briefly explain your reasoning.

[9] Keep in mind this question was asked in 2016 when Barack Obama was president.

D. Of the three issues, which one *most closely approximates* polarization? (circle one)

- ❑ Sending back children brought to U.S. illegally
- ❑ Legalization of marijuana
- ❑ Presidential approval

Briefly explain your reasoning.

E. Print the bar chart of the variable you chose in part D.

6. (Dataset: NES. Variables: cong_approve, cong_incumb_approve, nesw.) Pedantic pontificator claims he has discovered how voters evaluate the performance of House incumbents: "I call it my 'guilt by association' theory. When voters disapprove of the way Congress has been handling its job, they transfer that negative evaluation to their House incumbent. My theory is eminently plausible and surely correct. The distribution of opinions about House incumbents will be very similar to the distribution of opinions about the whole Congress."

The NES dataset contains cong_approve, which gauges respondent approval or disapproval of "the way the U.S. Congress has been handling its job." The dataset also has cong_incumb_approve, which measures approval or disapproval of the way each respondent's House incumbent "has been handling his or her job."

A. To test pedantic pontificator's theory, perform a frequencies analysis of cong_approve and cong_incumb_approve. Obtain bar charts with percentages. Refer to the Valid Percent column of the frequency distributions. In the table that follows, write the appropriate percentages next to each question mark.

	U.S. Congress	My House Incumbent
Approve strongly	?	?
Approve	?	?
Disapprove	?	?
Disapprove strongly	?	?
Total	100.00%	100.00%

B. Consider the evidence. Does pedantic pontificator's theory appear to be correct or incorrect? (circle one)

correct incorrect

Explain your reasoning.

7. (Dataset: States. Variables: defexpen, state.) Here is some conventional political wisdom: Well-positioned members of Congress from a handful of states are successful in getting the federal government to spend revenue in their states—defense-related expenditures, for example. The typical state, by contrast, receives far fewer defense budget dollars.

A. Suppose you measured the amount of defense-related expenditures in each state. The conventional wisdom says that when you look at the amount of defense-related expenditures in the United States, a few states would have a high amount of defense spending. Most states, however, would have lower values on this variable.

If the conventional wisdom is correct, the distribution of defense-related expenditures will have (circle one)

a negative skew. no skew. a positive skew.

If the conventional wisdom is correct, the *mean* of defense-related expenditures will be (circle one)

lower than its median. the same as its median. higher than its median.

B. The States dataset contains the variable defexpen, defense expenditures per capita for each of the fifty states. Perform a frequencies analysis of defexpen. In Statistics, obtain the mean and median, as well as skewness. (You do not need to obtain the mode for this exercise.) In the main Frequencies window, uncheck the "Display frequency tables" box. In Charts, request a histogram. Examine the results. Examine the histogram. Record the mean, median, and skewness next to the question marks in the table that follows.

Statistics for defexpen Variable	
Mean	?
Median	?
Skewness	?

C. Which is the better measure of central tendency? (circle one)

mean median

Briefly explain your answer.

_____

_____

_____

_____

_____

D. Print the histogram you produced in part B.

E. Based on your analysis, would you say that the conventional wisdom is accurate or inaccurate? (check one)

- ❑ The conventional wisdom is accurate.
- ❑ The conventional wisdom is inaccurate.

F. Use the Analyze ▶ Reports ▶ Case Summaries procedure to obtain a ranked list of states, from lowest per capita defense spending to highest per capita defense spending. The state with the lowest per capita defense spending is _____, with $_____ per capita. The state with the highest per capita defense spending is _____, with $_____ per capita.

8. (Dataset: States. Variables: blackpct_2016, hispanicpct_2016.) Two demographers are arguing over how best to describe the racial and ethnic composition of the "typical" state.

Demographer 1: "The typical state is 8.25 percent black and 8.20 percent Hispanic."

Demographer 2: "The typical state is 10.61 percent black and 11.26 percent Hispanic."

A. Run frequencies for blackpct_2016 (the percentage of each state's population that is African American) and hispanicpct_2016 (the percentage of each state's population that is Hispanic). Click the Statistics button to request the mean and median, as well as skewness. (You do not need to obtain the mode for this exercise.) In Charts, obtain histograms. In the main Frequencies window, uncheck the "Display frequency tables" box. Record the appropriate statistics for each variable in the table that follows.

	blackpct_2016	hispanicpct_2016
Mean	?	?
Median	?	?
Skewness	?	?

B. Based on your analysis, which demographer is more accurate? (circle one)

Demographer 1 Demographer 2

Write a few sentences explaining your reasoning.

_____

_____

_____

_____

C. Use the Analyze ▶ Reports ▶ Case Summaries procedure to obtain information on the percentage of Hispanics in the fifty states.

Which five states have the *lowest percentages* of Hispanics?

1. _____
2. _____
3. _____
4. _____
5. _____

Which five states have the *highest percentages* of Hispanics?

1. _____
2. _____
3. _____
4. _____
5. _____