2

DESCRIPTIVE STATISTICS

LEARNING OBJECTIVES

In this chapter you will learn to:

- Identify a variable's level of measurement
- Describe nominal-level variables using tables and figures
- Describe ordinal-level variables and evaluate their dispersion
- Describe interval-level variables with descriptive statistics and figures
- Sort datasets to obtain case-level information

EXCEL USAGE

- *Excel Companion* files: World Workbook
- File ▶ Open
- Insert ▶ PivotTable (with additional use of PivotTable options)
- Insert ▶ Recommended Charts ▶ Column (to make a *bar* chart)
- Data Analysis ToolPak: Descriptive Statistics, Histogram
- Data ▶ Sort

Analyzing descriptive statistics is the most basic—and sometimes the most informative—form of analysis you will do. Descriptive statistics communicate important attributes of a variable such as its typical value (central tendency) and its spread (degree of dispersion or variation).

How you describe a variable depends on its level of measurement. The higher the level of measurement, the more detailed the description. For nominal-level variables, for example, the *mode*, the most common value of the variable, is the statistic that describes central tendency. 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-sized 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. Describing a variable's degree of dispersion or variation often requires informed judgment.[1]

In this chapter you will use Excel functions and data analysis tools to obtain appropriate descriptive statistics for variables with different levels of measurement. You will also learn how to use descriptive statistics to better understand variables and communicate their most important features. With the correct prompts, Excel can produce valuable graphic support—bar charts and (for interval variables) histograms. These tools are essential for distilling useful information from large datasets. For smaller datasets with aggregated units, such as the States Workbook's dataset and the World Workbook's dataset, Excel allows you to sort observations based on the values of a variable to report case-level information about a variable that you find especially interesting.

2.1 IDENTIFYING LEVELS OF MEASUREMENT

Suppose you were hired by a telephone-polling firm to interview people. Your job is to find out and record three characteristics of each person you interview: their age, political ideology, and birthplace. These variables—age, political ideology, and birthplace—are three pieces of information about people that vary. You might describe a respondent this way: "The respondent is 22 years old, is ideologically moderate, and was born in Kansas." This would be a good thumbnail description, easily interpreted by another person. These three pieces of information about a person have different levels of measurement, which shapes how we describe their variation among people.

READING IN *ESSENTIALS*

Read Chapter 2, pages 34–55, in the sixth edition of *The Essentials of Political Analysis* to learn how variables are measured and described in political science.

Some variables have qualitative values. For example, when we ask someone where they were born, their response is a place, such as Kansas, Atlanta, or Mexico. Everyone was born somewhere and a variable like birthplace simply identifies the place. Birthplace is a nominal-level variable. Similarly, when we ask someone their political ideology, their response is a phrase like "ideologically moderate," which expresses the value of this varying characteristic in words. Political ideology, like birthplace, is qualitative information, but its values can be ordered, making it a variable measured at the ordinal level. Some people are ideologically moderate, some are extremely liberal, and others are extremely conservative. We could ask people to identify their political ideology along a spectrum that runs from extremely liberal on one side, to moderate in the middle, to extremely conservative on the other side.

Level of measurement: the precision with which a variable is measured. Some values are qualitative (nominal or ordinal level), while others are quantitative (interval level).

Some variables, like someone's age in years, provide quantitative information. Variables measured at the interval level provide precise, numerical information about the observations. We can describe the central tendency and dispersion of any variable, but the higher the variable's **level of measurement**, the larger our toolkit for describing it. When a variable's values are meaningful numbers, we analyze them with math in ways that are not possible when a variable's values are words.

How you describe a variable depends on the variable's level of measurement. When a variable records qualitative information about the observations, the methods available to describe it are relatively limited. When a variable's values quantify characteristics of the units of analysis with numbers, there are more tools available to describe the variable. These and other points are best understood by working through some guided examples.

2.2 DESCRIBING NOMINAL VARIABLES

For this and the next few analyses, you will use the World Workbook's dataset. Open the World Workbook's dataset by double-clicking the world.xlsx file or, if you already have Excel running, select File ▶ Open and locate world.xlsx.

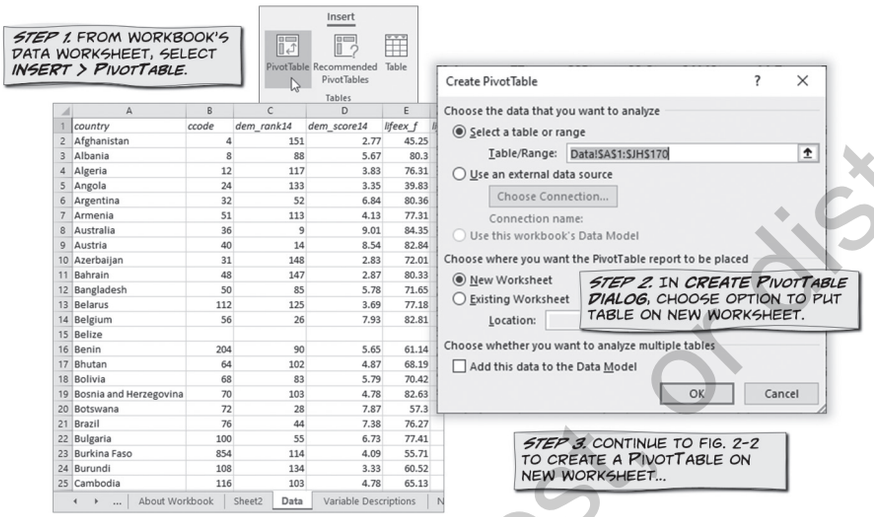
[1]In this chapter we use the terms *dispersion*, *variation*, and *spread* interchangeably.

2.2.1 Frequency Distribution Tables

We can describe the distribution of a nominal-level variable in the World Workbook's dataset, regionun, which classifies countries in the world by region, with a **frequency distribution table**. This table will list world regions by row and show us how many countries are in each region. To create a frequency distribution table with Excel, start on the worksheet that contains the variable and then select Insert ▶ PivotTable to summon the Create PivotTable dialog (see Figure 2-1).[2]

Frequency distribution table: shows the relative frequency of a nominal- or ordinal-level variable's values in percentages or proportions; may also show counts.

FIGURE 2-1 ⬢ Inserting a Pivot Table into a New Worksheet



	A	B	C	D	E
1	country	ccode	dem_rank14	dem_score14	lifeex_f
2	Afghanistan	4	151	2.77	45.25
3	Albania	8	88	5.67	80.3
4	Algeria	12	117	3.83	76.31
5	Angola	24	133	3.35	39.83
6	Argentina	32	52	6.84	80.36
7	Armenia	51	113	4.13	77.31
8	Australia	36	9	9.01	84.35
9	Austria	40	14	8.54	82.84
10	Azerbaijan	31	148	2.83	72.01
11	Bahrain	48	147	2.87	80.33
12	Bangladesh	50	85	5.78	71.65
13	Belarus	112	125	3.69	77.18
14	Belgium	56	26	7.93	82.81
15	Belize				
16	Benin	204	90	5.65	61.14
17	Bhutan	64	102	4.87	68.19
18	Bolivia	68	83	5.79	70.42
19	Bosnia and Herzegovina	70	103	4.78	82.63
20	Botswana	72	28	7.87	57.3
21	Brazil	76	44	7.38	76.27
22	Bulgaria	100	55	6.73	77.41
23	Burkina Faso	854	114	4.09	55.71
24	Burundi	108	134	3.33	60.52
25	Cambodia	116	103	4.78	65.13

By default, the entire data worksheet should be selected to analyze, which is what we want at this stage. Select the option to place the **PivotTable** report in a new worksheet to keep data analysis separate from the dataset. If you have an existing worksheet where you save the results of your analysis, select it by location instead of reporting to a new worksheet. Click OK and Excel switches to a new worksheet, ready to create a pivot table (see Figure 2-2).

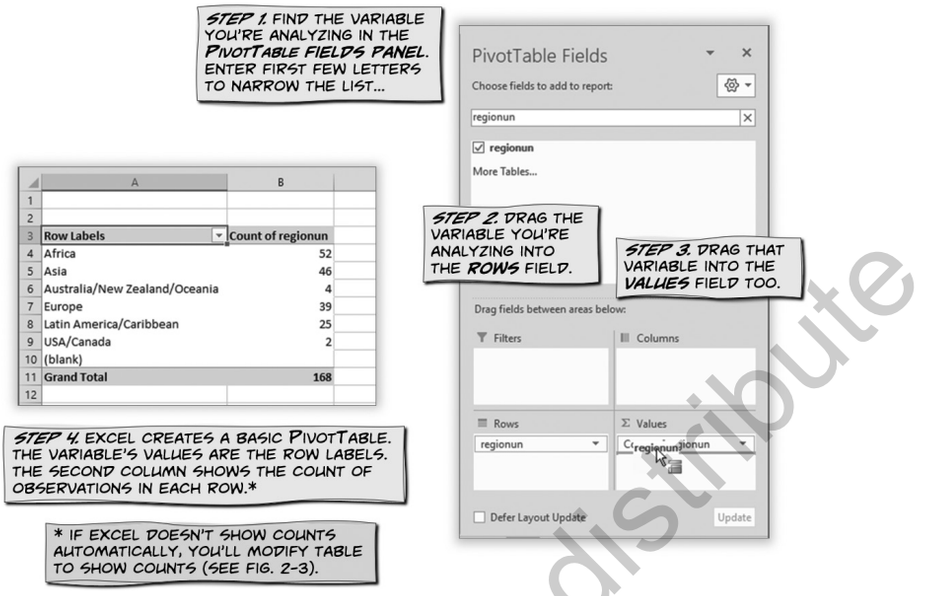
PivotTable: a versatile Excel procedure for summarizing records in table format.

You can do a lot with pivot tables. Here, we're creating a frequency distribution table of the observed values of the regionun variable. Begin by checking "regionun" as the field to add to the report (see Figure 2-2). If you can't find this variable quickly, click the gear icon and select the option to sort the variable list from A to Z or start typing the variable name to focus the list of options. After you select regionun, you'll see all the different observed values of the variable listed in alphabetical order (by default). The variable's values define the pivot table's rows.

After choosing the variable that defines the row labels, we'll tell Excel how to calculate the values we want to see in a frequency distribution table. We want this table to communicate how many countries are located in the different regions of the world. Drag "regionun" from the variable list into the Values section in the bottom-right corner of the PivotTable Fields panel. Excel adds a column to the pivot table that reports the number of countries in each region as well as a grand total.

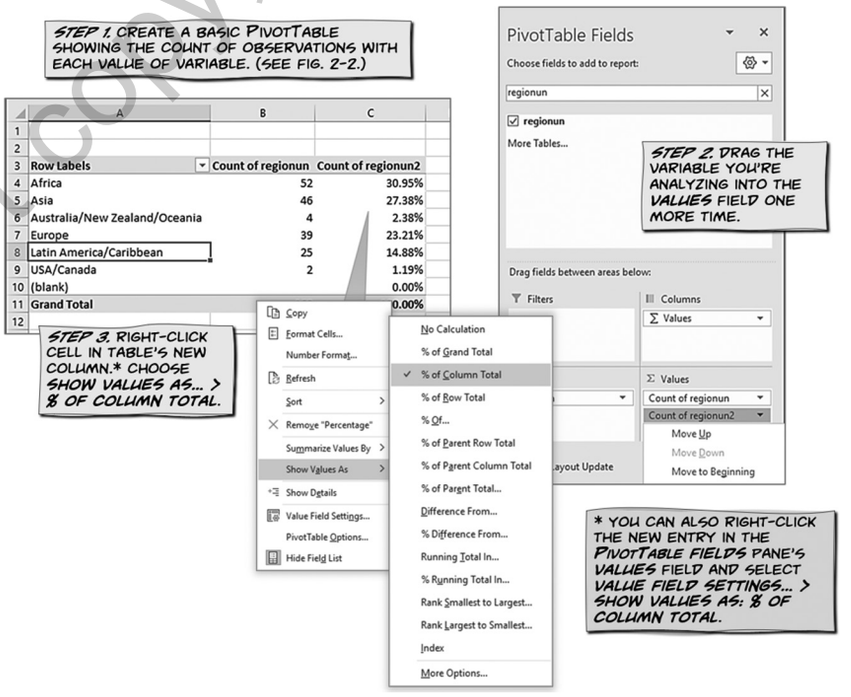
[2]Truth be told, there are several ways to generate most descriptive statistics with Excel, including a frequency table. The pivot table is Excel's most versatile tool for data analysis, so we're focusing on it first.

FIGURE 2-2 ⬢ Using a Pivot Table to Show a Frequency Distribution



The counts of regionun are useful, but a frequency distribution table should also report percentages. To add a column for percentages, drag "regionun" from the variable list into the Values section of the worksheet a second time (see Figure 2-3). We don't want a duplicate column of counts, so right-click the new column of the pivot table and select the "Value Field Settings" option from the pop-up menu. In the Value Field Setting dialog, click the Show Values As tab, and then select "% of Grand Total" from the menu of options for showing the variable's values.

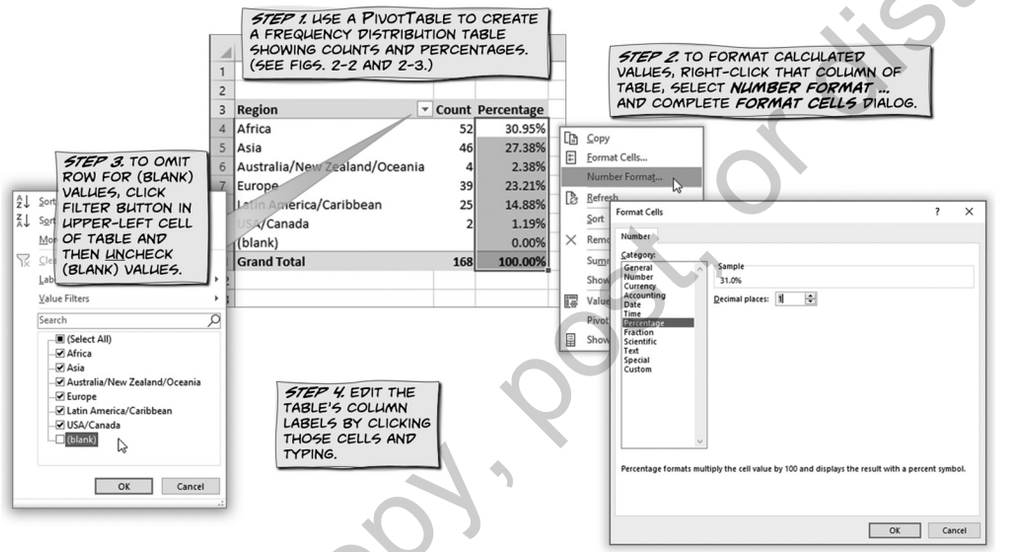
FIGURE 2-3 ⬢ Adding a Percentages Column to a Frequency Distribution Table



After completing the steps illustrated in Figure 2-3, the main elements of the frequency distribution table of the regionun variable are complete. All possible values of the variable are listed by row; one column displays the count of countries in each region and another column displays the percentages.

After you generate a pivot table, you'll usually want to edit the table to make it clearer and more concise. To communicate descriptive information more clearly, we can filter out the cases where the variable's value is blank, edit the column labels, and adjust the number of decimal places displayed for percentages. To filter out the row labeled "(blank)," click the down-arrow button on the right side of the Row Labels cell, uncheck the blank entries, and then click OK. To edit a column label, simply click on the label and type the new text. You can adjust the number of decimal places displayed by dragging the cursor over the cells with numbers to select them, right-clicking the selection, choosing the Number Format option, and then adjusting the number of decimal places. Excel generally displays too many decimal places by default, making numeric results appear more complicated than they really are. One or two decimal places generally suffice and fewer digits make tables of numbers clearer.[3] Excel makes the requested alterations to the frequency distribution table (Figure 2-4).

FIGURE 2-4 ⬢ Editing a Frequency Distribution Table



Sometimes displaying row labels in alphabetical order is the best option, but in some cases you'll want to change the order of rows. To order the rows arbitrarily (i.e., not alphabetically or reverse alphabetically), right-click a row label, like "USA/Canada," select Move from the pop-up menu, and then inform Excel whether you want to move the selected row label to the beginning (first row) of the table, up one row, down one row, or to the end (last row before grand total) of the table. With a bit of fine-tuning, you'll get the frequency distribution table in the right order.

Region	Percentage	Count
Africa	31.0%	52
Asia	27.4%	46
Australia/New Zealand/Oceania	2.4%	4
Europe	23.2%	39
Latin America/Caribbean	14.9%	25
USA/Canada	1.2%	2
Grand Total	**100.0%**	**168**

[3]If you're working with very small numbers, you should make sure your tables report at least one nonzero number in every cell.

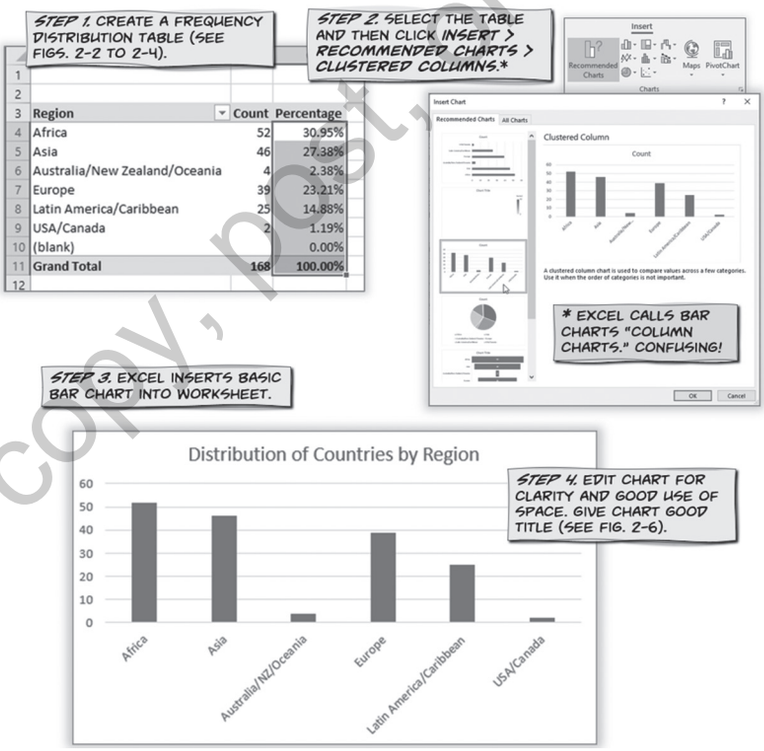
The "Count" column shows the number of countries in each region of the world. We can consult this table to identify the region with the most countries, Africa, and see that there aren't many countries in some regions, such as the USA/Canada region.

2.2.2 Bar Charts

Bar chart: a visual depiction of the relative distribution of a nominal- or ordinal-level variable's values. Excel calls bar charts "column charts."

A visual representation of the regionun variable can help describe its typical values and dispersion. Often, researchers use tables and graphics in tandem to communicate the results of their data analysis. A **bar chart** pairs nicely with a frequency distribution table to describe the values of a nominal-level variable. A bar chart will show the count of countries in each region with bars of varying heights. For this simple display, the charting tools associated with Excel's pivot tables are overly complicated.[4] To create a bar chart for the frequency distribution table, select the regions and counts from the pivot table (do not select the grand total row), copy the selection, and paste it below the pivot table. With the copied subsection of the frequency distribution table selected, click Insert ▶ Recommended Charts to summon the Insert Chart dialog (see Figure 2-5). We're going to select a column chart, but you'll notice there are many ways to show how many countries are in different regions. When you browse chart options, Excel will preview the results to help you select the right chart. Select a column-style bar chart and click OK. To create the final bar chart shown in Figure 2-5, we abbreviated "New Zealand" to "NZ" and gave the chart a descriptive title.

FIGURE 2-5 ⬢ Creating a Bar Chart



2.2.3 Central Tendency and Dispersion

Scroll between the frequency distribution table and the bar chart, which depicts the regionun variable in graphic form. What is the mode, the region with the most countries? For nominal variables, the answer to this question is (almost) always an easy call: Simply find the value with the largest count,

[4]Excel's Pivot Chart tools make it difficult to display only select columns of a pivot table.

highest percentage of responses, and tallest bar. Africa is the modal region. When it comes to describing the central tendency of nominal-level variables like this, our toolkit is limited to identifying the variable's mode.

Does the region variable have little dispersion or a lot of dispersion? Again, study the frequency distribution table and the bar chart. Apply the following rule that applies to any variable at any level of measurement: *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 countries located in Africa, or are there many countries in each region? There are a lot of countries in four of these six regions, but there are few countries in two of them. Countries aren't evenly dispersed by region, and they also are not clustered in just one or two regions: medium-level dispersion. When looking at the bar chart of a nominal-level variable like the one in Figure 2-5, ignore the order of the bars. If we relocated the USA/Canada bar to one of the center positions, the distribution would look more spread out, but remember that the order of nominal-level values is arbitrary.

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.

In other instances, however, you may find 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, to identify their main source of political news: television, websites, or friends and family. Among your own acquaintances, you probably know a number of people who fit into each category. So even if one category, such as "television," is the mode, many people will say websites or friends and family. In this instance, the amount of dispersion in a variable—its degree of spread—is essential to understanding and describing it.

2.3 DESCRIBING ORDINAL VARIABLES

In this section, you will analyze and describe two ordinal-level variables in the World Workbook's dataset, one of which has little variation and the other of which is more spread out. One of these ordinal-level variables records the level of Internet freedom in countries around the world; the other, the level of civil war in countries around the world. Both variables are recorded on 3-point ordinal scales.

2.3.1 High Dispersion Example

To create a frequency distribution table of Internet freedom, we will use the same function we did to describe the nominal-level variable regionun. Start on the worksheet that contains the dataset and then select Insert ▶ PivotTable to summon the Create PivotTable dialog (see Figure 2-1). Make sure you generate frequency distribution tables for these variables on a different worksheet than the data.

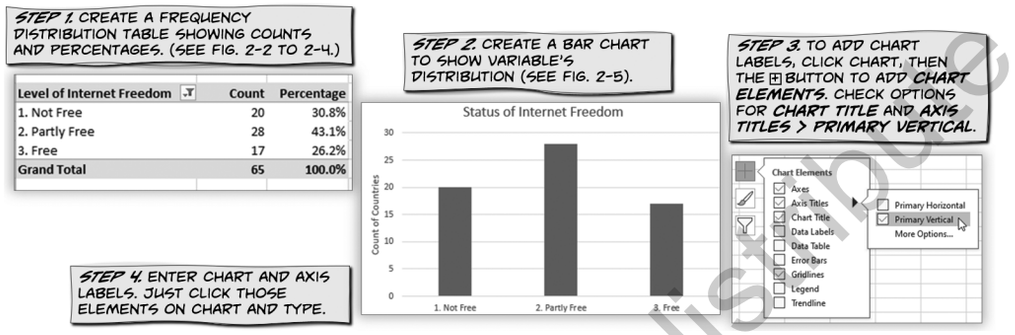
To create a table to describe the level of Internet freedom around the world, scroll through the variable list until you find the variable internet_status and check it. Excel uses this variable's three values to define the table's rows: (1) not free, (2) partly free, and (3) free. Refer to Figure 2-2 to create the basic elements of the frequency distribution table for internet_status and Figure 2-3 to edit the table for clarity. This measure of Internet freedom (from a 2016 Freedom House report[5]) does not cover all countries in the world, but we can filter out blank entries and focus on the available data.

Figure 2-6 shows an Excel worksheet with both a frequency distribution table and a bar chart for the internet_status variable. We made one additional edit to the bar chart. Click anywhere on the chart and three small squares appear to its right. Clicking the top square, which looks like a plus sign, gives

[5]See Freedom House, "Freedom in the World 2016," https://freedomhouse.org/sites/default/files/FH_FITW_Report_2016.pdf.

you a menu to show/hide various chart elements, such as the axis titles for the vertical and horizontal axes. We've added a vertical axis title and edited it to read "Count of Countries." This variable's values are numbered so they display in a logical order when arranged alphabetically, but whether they're displayed in ascending or descending levels of Internet freedom is arbitrary (the middle category stays the same).

FIGURE 2-6 ⬢ Describing an Ordinal-Level Variable with High Dispersion



Because internet_status is an ordinal variable, you can use both its mode and its median to describe central tendency. Its mode, clearly enough, is "2. Partly free," with 43.1 percent of the countries. What about the median? *The median for any ordinal (or interval) variable is the category below which 50 percent of the cases lie.* Is the first category, "1. Not free," the median? No, 30.8 percent of countries fall into this category, less than half of the cases. How about the next higher category? Yes, the median occurs in the "2. Partly free" category (cumulative percentage, 73.9). This variable has the same median and mode values.

Now consider the question of whether internet_status has a high or low degree of dispersion. If internet_status has a high level of variation, then the percentage of countries in each category would be roughly equal, with one-third of cases (33.33 percent) in each of the categories. If internet_status has no dispersion, then all the cases would fall into one value. That is, one value would represent 100 percent of the countries, and each of the other categories would represent 0 percent. Which of these two scenarios comes closest to describing the distribution of internet_status? Examining the descriptive statistics, we see that the percentages in each category aren't exactly equal, but they're in the same vicinity: 30.8, 43.1, and 26.2 percent. The middle bar, which represents countries where the Internet is partly free, is the tallest but this variable has a high level of dispersion.

2.3.2 Low Dispersion Example

Let's contrast the distribution of internet_status with the distribution of the civil_war variable's values. The civil_war variable, which also uses a 3-point scale, identifies the intensity of domestic conflict in countries around the world: major, minor, or no civil war. You can produce a frequency distribution table and bar chart for the civil_war variable the same way we did to generate descriptive statistics for the internet_status variable. Review Section 2.3.1 as necessary to do this analysis.

Civil war status	Percentage	Count
Major civil war	3.1%	5
Minor civil war	10.7%	17
No civil war	86.2%	137
Grand Total	**100.0%**	**159**

FIGURE 2-7 ⬢ Bar Chart of an Ordinal Variable with Low Dispersion

The frequency distribution table and bar chart of the civil_war variable communicate important information. What value is most typical? Clearly, the modal value is "no civil war," with 82.2 percent of countries. The bar for "no civil war" towers over the bars for "major" and "minor" civil wars. With the vast majority of cases, "no civil war" is both the mode and median value. What about this variable's dispersion? Remember, if civil_war has a high level of variation, then roughly one-third of countries (33.33 percent) would fall into each category. If civil_war has no dispersion, then all the cases would fall into one value. The civil war status of countries falls mostly into the "no" category; this variable has very low dispersion.

2.4 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.

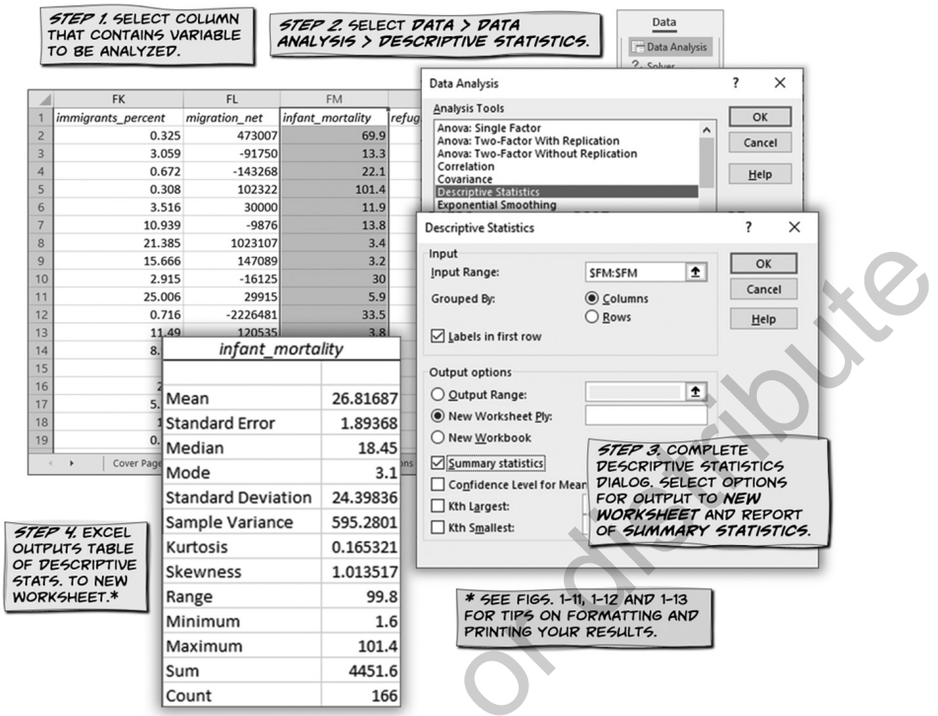
Because interval variables have the most precision, they can be described more completely than can nominal or ordinal variables. We have a relatively large toolkit available for describing variables measured at the interval level. For any interval-level variable, you can report its mode, median, and 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*.

2.4.1 Descriptive Statistics with the Data Analysis ToolPak

A step-by-step analysis of a World Workbook variable, infant_mortality, will clarify these general observations. This variable quantifies the number of infants who die before age 1 year per 1,000 births in countries around the world. A country's infant mortality rate is an important and widely used health indicator.

You can use Excel's descriptive statistics tool to generate a table of descriptive statistics for an interval- level variable. If you haven't already installed the Analysis ToolPak add-in, you'll need to install it now to use the descriptive statistics tool (see Section 1.2). To access the descriptive statistics tool, select Data ► Data Analysis ► Descriptive Statistics and you should see the Descriptive Statistics dialog (see Figure 2-8). For the input range field, select the worksheet column with the infant_mortality variable. This variable has a label, the variable name, in the first row. Request summary statistics. Make sure the output goes to a different worksheet than the data and click OK.

FIGURE 2-8 ⬢ Generating Descriptive Statistics for an Interval-Level Variable



Excel generates a table of descriptive statistics for the infant_mortality variable. For clarity, we've limited the numbers—except for "count," which is a whole number—to two decimal places. The table of descriptive statistics offers several measures of central tendency. The mean infant mortality rate among 166 countries for which data are available is 26.82 deaths before age 1 year per 1,000 live births. We'll skip over "standard error" for now but return to it in Chapter 8. The median infant mortality rate is 18.45; 50 percent of countries have a higher infant mortality rate than this and 50 percent have a lower rate. When you're analyzing a variable with continuous values, mode is usually a meaningless statistic because few observations will have the exact same value and only due to rounding.

Excel's descriptive statistics tools also generate multiple statistics to describe the infant_mortality variable's dispersion. One measure of dispersion, range, is the difference between the variable's maximum value and its minimum value (101.40 − 1.60 = 99.80). The variable's standard deviation, 24.40, tells us how much the values of infant_mortality observed in countries around the world typically deviate from the variable's mean value. The sample variance statistic, 595.28, equals the standard deviation raised to the second power ($24.40^2 = 595.28$).

Descriptive Statistics for infant_mortality	
Mean	26.82
Standard Error	1.89
Median	18.45
Mode	3.10
Standard Deviation	24.40
Sample Variance	595.28
Kurtosis	0.17
Skewness	1.01
Range	99.80
Minimum	1.60
Maximum	101.40
Sum	4451.60
Count	166

In addition, Excel's descriptive statistics table reports values for skewness and a statistic called kurtosis. **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. When a distribution is perfectly symmetrical—no skew—it has skewness equal to 0. If the distribution has a larger and longer right-hand tail—positive skew—then skewness will be a positive number. A more pronounced left-hand tail, logically enough, returns a negative number for skewness. 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.* We won't do much with the **kurtosis**, which measures whether the tails of a distribution are heavier (positive kurtosis values) or lighter (negative kurtosis values) than normal.

Skewness: a statistical measure of a distribution's symmetry. If positive, the mean is greater than the median and the distribution has a longer right tail. If negative, the mean is less than the median and the left tail is longer.

Kurtosis: a statistical measure of a distribution's peakedness. If greater than 3, it's more peaked than a bell curve; if less than 3, it's less peaked than a bell curve.

For the infant_mortality variable, the skewness statistic is positive (1.01). This suggests that the distribution of values has a longer right-hand tail, which means infant_mortality rates have more dispersion above the mean than below the mean. Note also that the mean (26.82) is higher than the median (18.45), a situation that often—although not always—indicates a positive skew.[6] 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.[7]

2.4.2 Histograms

A chart of infant_mortality rates observed around the world helps communicate the essential features of this variable. 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 should generate 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 count) of cases that fall into each category. A histogram is similar to a bar chart, but instead of displaying each of the variable's unique values, it uses value ranges (called bins) to define categories, resulting in a compact display. Histograms are sometimes more readable and elegant than bar charts. 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.)

Histogram: a chart that shows the distribution of an interval-level variable's values. Each vertical bar represents a binned range of values.

To create a histogram of infant_mortality rates observed around the world, one first defines the value intervals for the bins. Each observed value should fall into exactly one bin. You don't want to overcomplicate the histogram with too many bins, but you also want the histogram to describe the distribution of values accurately. Since each bin covers 10 units, we'll need 11 bins to cover the full range of values and the histogram should have the right amount of detail. Next to the table of descriptive statistics, enter 10, 20, 30 . . . 110 in a column with heading "Bins." These values define the upper boundary of each bin; the first bin represents countries with infant mortality rates up to 10, the next bin captures countries over 10 and up to 20, and so on, with the last bin representing observations over 100 and up to 110 (the maximum observed value is 101.40 so there is no need for a bin above 100–110).[8] Once you've defined the bins, select Data ► Data Analysis ► Histogram to summon the Histogram tool dialog (see Figure 2-9).[9]

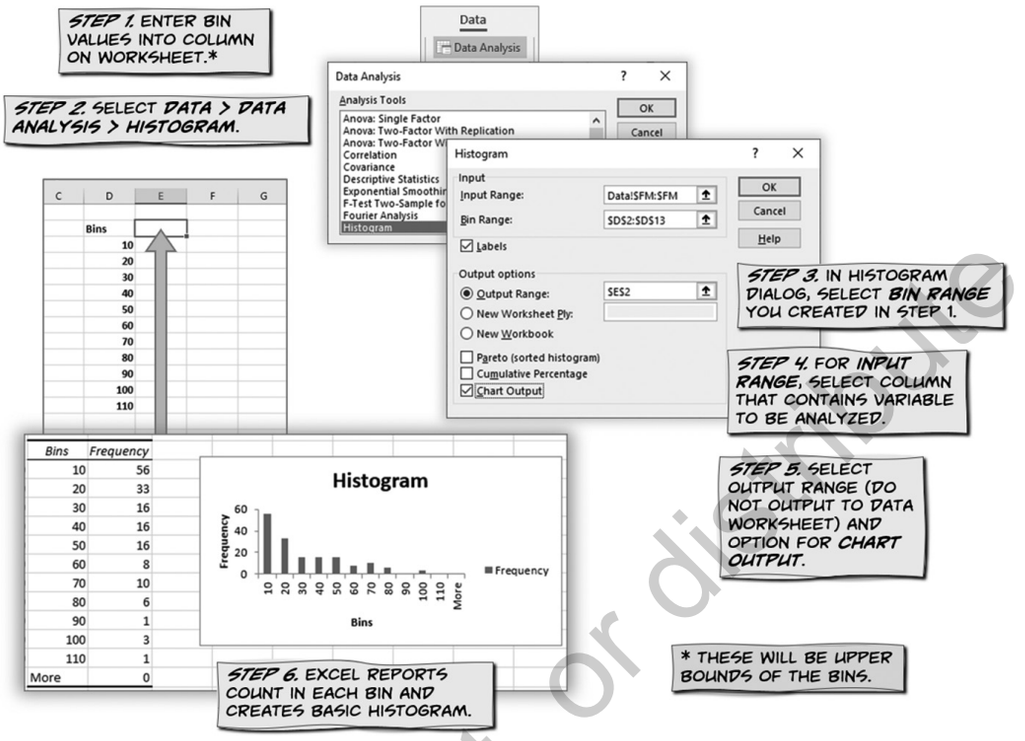
[6]Paul T. von Hippel, "Mean, Median, and Skew: Correcting a Textbook Rule," *Journal of Statistics Education* 13, no. 2 (2005). "Many textbooks teach a rule of thumb stating that the mean is right of the median under right skew, and left of the median under left skew. This rule fails with surprising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

[7]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.

[8]The bin includes the upper boundary.

[9]You can also use Insert Charts to create a histogram; this method is quicker than the data analysis tool, but it doesn't give you any control over the bin intervals.

FIGURE 2-9 ⬢ Creating a Histogram to Describe an Interval-Level Variable



The input range for the histogram is the column of the data worksheet with the infant_mortality variable. To select these data, you'll need to navigate to the data worksheet, select that column, and then return to the separate page you've created for the infant_mortality variable's descriptive statistics and histogram. The Histogram dialog should then include the name of the data worksheet in the input range—for example, "Data!$FM:$FM" for column FM of the Data worksheet. Select the field that defines the upper limits of the bins for the bin range, including the heading "Bins." The input values and bin values both have labels, so that box gets checked. For Output options, select the cell to the right of the heading "Bins" for the Output Range and check the option for Chart Output. Once you've completed the Histogram tool dialog (see Figure 2-9), click OK to generate the histogram.

You should think of an Excel chart like the one seen in Figure 2-9 as a rough draft of the chart you're making. Consider, first, whether bin intervals are set in a way that helps communicate the essential features of the variable. If you have too many bins, not enough bins, or need to modify the intervals, edit the column of values you made and repeat the steps outlined in Figure 2-9 (the chart won't automatically update but your selections are saved in the Histogram tool dialog and easily revised). Once the basic structure is right, edit chart elements and format the histogram so it communicates the essential features of the variable as clearly as possible. Here, even the rough cut of the infant_mortality histogram shows that most countries fall into the 0–10 range and the distribution has a long right-side tail, with infant mortality rates in a few countries multiple times higher than the rates in most of the world.

A CLOSER LOOK: Editing Charts with Purpose

Excel makes it easy to modify the content and appearance of any chart it creates. To edit most chart elements, you simply double-click on the element you want to edit, and Excel lets you edit it directly or displays your editing options. When you click on a chart element, Excel also adds Chart Tools to the ribbon. The Design and Format options allow you to quickly change color schemes and the overall chart appearance.

The variety of chart editing options and their ease of use is great, but you should keep some design principles in mind so these options help and don't hinder your ability to communicate the results of your analysis with graphics. If you're going to use an Excel chart in a paper or presentation, try 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.

You should view the area of your chart the way an artist looks at a canvas. Make effective use of the space. Your chart should be informative and clear without being cluttered. If your chart shows a few dots in a sea of white space, you've wasted your canvas. If you have to explain your chart so others understand it, you haven't made it clear enough.

When it comes to colors and overall appearance, consider your audience. Academic audiences tend to be very conservative when it comes to graphic design. Don't try to make your charts fun or pretty. If you're creating a chart for a presentation, the audience may appreciate some color and attention-grabbing design, but you don't want your charts to become distractions or cause people to take you less seriously. If you're unsure about your audience's expectations, look at the charts and graphics in academic articles or presentations.

2.5 CASE-LEVEL INFORMATION

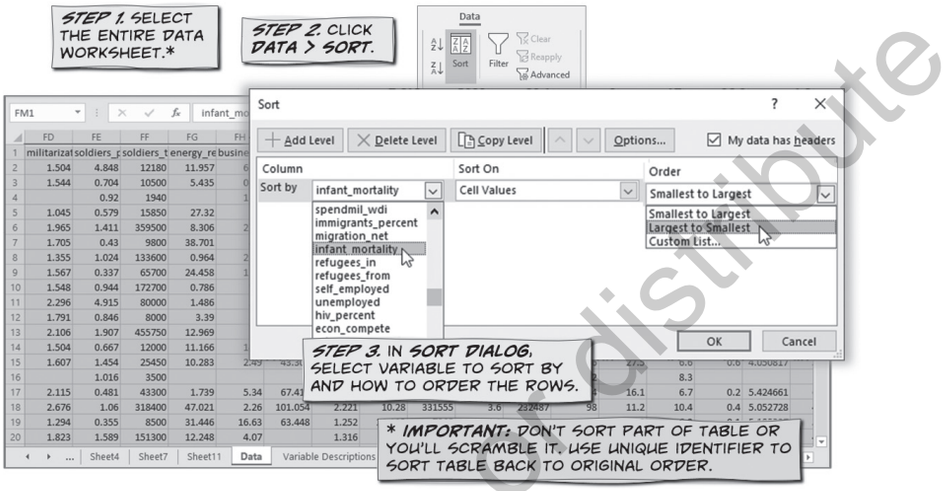
When you analyze a large survey dataset, you generally are not interested in how respondent X or respondent Y answered a particular question; they're just some random people who happened to participate in the survey. Sometimes, however, you gather data on particular cases because the cases are themselves inherently important.

When you work with the States dataset or the World dataset, you may want to describe cases beyond the relative anonymity of summary statistics and find out where particular states or countries "are" on an interesting variable. Excel makes it easy to sort cases based on the value of a variable in order to identify and report case-level information.

Suppose you are interested in identifying countries that have the highest/lowest infant mortality rates in the world. As we saw in the preceding section, countries vary considerably on this important health measure. To sort countries by ascending or descending infant mortality rates, highlight the entire data worksheet (you can select all rows and columns by pressing Ctrl-A). Next, select Data ▶ Sort, which calls up the Sort dialog (see Figure 2-10). Scroll down the variables listed in the "Sort by" field until you find infant_mortality and select it. We'll sort cases from lowest to highest values of infant_mortality first, so you can keep the other Sort dialog settings. Click OK and Excel sorts the rows of the worksheet in the desired order. It is not immediately clear which countries have the lowest infant mortality rates

because the column containing country names is not immediately adjacent to the infant_mortality variable (the country names are in column A). To see the country names and infant mortality rates side by side, select all the columns between "country" and "infant_mortality," right-click these columns, and then select the option to "Hide" them.

FIGURE 2-10 ⬢ Sorting Cases by Variable Values



To identify countries with the highest infant mortality rates, follow the steps just discussed and outlined in Figure 2-10, except change the sort order from "smallest to largest" to "largest to smallest." We can copy the relevant results from the sorted Data worksheet to identify countries with the lowest and highest infant mortality rates in the world.

Lowest Infant Mortality Rates		Highest Infant Mortality Rates	
Country	**Infant Mortality Rate**	**Country**	**Infant Mortality Rate**
Iceland	1.6	Angola	101.4
Luxembourg	1.6	Central African Republic	96.1
Finland	2.1	Sierra Leone	93.8
Japan	2.1	Somalia	90.1
Singapore	2.2	Chad	88.4

This case-level information about infant mortality rates in different countries does not tell us why some countries have higher infant mortality rates than others do, but this kind of information can supplement summary tables and figures and help us develop theories. Excel's Data ▸ Sort routine is great because it not only enables you to sort based on the values of one variable in ascending/descending order as we have shown here, but it also allows you to sort by multiple criteria. You could, for example, sort countries of the world by civil war status first (major, minor, or none, the ordinal-level variable we described earlier in the chapter) and then by infant mortality rate within each of the civil war statuses.

CHAPTER 2 EXERCISES

Name: _____ Date: _____

E-mail: _____ Section: _____

1. How you analyze a variable depends on its level of measurement. In order to apply the right methods, you must be able to identify a variable's level of measurement.[10]
   A. The States Workbook's dataset includes a variable named min_wage, which is the minimum wage in each state in dollars and cents. What's the level of measurement of the min_wage variable? (select one)

      Nominal Ordinal Interval

   B. The World Workbook's dataset includes a variable named frac_eth3, which records the level of ethnic fractionalization in countries as low, medium, or high. What's the level of measurement of the frac_eth3 variable? (select one)

      Nominal Ordinal Interval

2. Practice identifying the level of measurement of variables by completing the following table.[11]

Workbook	Variable	Level of Measurement (check one)	How Do You Know?
States	voter_id_law	❑ Nominal ❑ Ordinal ❑ Interval	
States	opioid_rx_rate	❑ Nominal ❑ Ordinal ❑ Interval	
World	gender_equal3	❑ Nominal ❑ Ordinal ❑ Interval	
World	typerel	❑ Nominal ❑ Ordinal ❑ Interval	

3. The General Social Survey (GSS) includes ten true-or-false questions to test 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).
   A. Complete a frequency distribution table of science_quiz values. Fill in the table that follows.[12]

Quiz Score	Frequency	Percent	Cumulative Percent
0	2		
1	9		
2	13		
3	38		
4	54		
5	66		
6	80		

(Continued)

[10]Section 2.1 tells you how to identify a variable's level of measurement.

[11]Subsequent chapters build on your ability to identify levels of measurement, so you must master this skill.

[12]See Section 2.2.1 for guidance on frequency distribution tables. Section 2.2.1 uses this type of table to describe a nominal-level variable but you can use it for this interval-level variable because it has only eleven unique values.

(Continued)

Quiz Score	Frequency	Percent	Cumulative Percent
7	78		
8	60		
9	45		
10	21		100.0%
Total	466	100.0%	

B. Create a bar chart of the distribution of science_quiz scores. Submit the bar chart with your answers. (Alternatively, you can create a histogram, but there is no need to group observations into different binned values of science_quiz values).[13]

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

Mean Median

D. Briefly explain your choice.

_____

_____

_____

_____

_____

_____

E. 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? (select one)

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

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

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

4. The GSS 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 following frequency distribution table provides the count of GSS respondents for each value of the attend variable. Complete the table by calculating the appropriate column percentages.[14]

Variable Value	Percentage	Count
Never		711
Less than once a year		168
Once a year		378
Several times a year		317
Once a month		199
Two to three times a month		249
Nearly every week		127
Every week		498
More than once a week		204
Total	100.0%	2,850

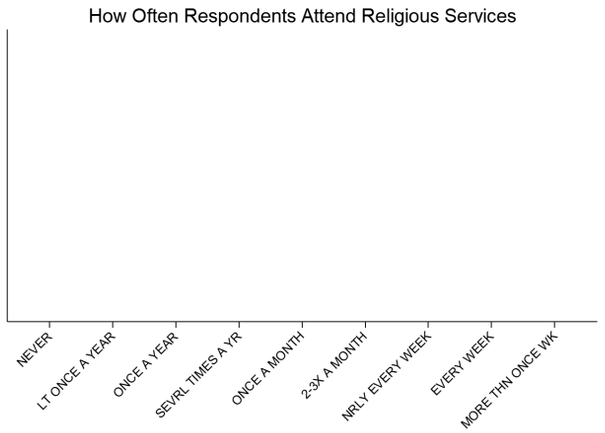
B. Based on your examination of the frequency distribution, the mode of attend is _____. The median of attend is _____.

C. In the following graphic shell, create a bar chart of how often Americans attend religious services based on the frequency distribution table for completed in part A.[15]

[13]Section 2.2.2 shows you how to create a bar chart. Section 2.2.2 applies this skill to a nominal-level variable but you can apply it to science_quiz because it has only eleven unique values. Section 2.4.2 covers histograms.

[14]We discuss frequency distribution tables in Section 2.2.1.

[15]See Section 2.2.2 for further discussion of bar charts.



D. Based on your examination of the frequency distribution and bar chart, you would conclude that attend has (select one)[16]

low dispersion. high dispersion.

5. In this exercise, you'll create some frequency distribution tables for variables in the World Workbook's dataset.[17] As you complete the following tables, pay attention to detail. Report the *proportion* of cases that fall into each category.[18] If you generate a table with percentages, convert those percentages to proportions. Exclude missing values from the table and your calculations.

A. Variable: rights_injud

Independence of Judiciary	Proportion	Count
Complete		
Some		
None		
Total	1.000	

B. Variable: rights_speech, freedom of speech

Freedom of Speech	Proportion	Count
Complete		
Some		
None		
Total	1.000	

6. You can create frequency distribution tables for variables measured at the nominal level as well as variables measured at the ordinal level. The tables look similar but you can add a column of cumulative percentages when you're working with an ordinal-level variable, although you can't add a column of cumulative percentages to the frequency distribution table of a nominal-level variable.[19] Why is that?

_____

_____

_____

7. Both bar charts and histograms are used to visually display the dispersion of a variable's values. Bar charts and histograms sometimes look very similar but there are important differences between them.[20] How are histograms different than bar charts? Why would you use a histogram to display the dispersion of an interval-level variable instead of a bar chart?

_____

_____

_____

[16]Be sure to read Section 2.3, which compares an ordinal-level variable with high dispersion to one with low dispersion.

[17]This skill is covered in Section 2.2.1 and applied to ordinal-level variables in Section 2.3.

[18]These variables are coded numerically. For rights_injud and rights_speech, 0 = none, 1 = some, and 2 = complete. For rights_wopol, 1 = none, 2 = some, and 3 = complete.

[19]To answer this question correctly, you need to understand the difference between nominal- and ordinal-level variables (covered in Section 2.1) and apply that understanding to table construction.

[20]We discuss histograms as an alternative to bar charts in Section 2.4.2.

8. Two political pundits have a debate about the reliability of death sentences in the United States. In this exercise, you'll analyze the deathpen_exonerations variable in the States Workbook's dataset to assess the claims made by these two pundits.

   Pundit 1: There's no such thing as a perfect trial. To err is human. Witnesses identify the wrong person in a lineup. Police think they've solved the case. It's unfortunate, but it happens all over the country. A lot of death sentences have been exonerated across the United States.

   Pundit 2: I agree that a lot of people have been mistakenly sentenced to death, but it hasn't happened all over the country. Exonerations are not widely dispersed across the country; they're concentrated in a small number of states.

   A. Apply the Descriptive Statistics tool in Excel's Data Analysis ToolPak to the deathpen_exonerations variable.[21] Fill in the table that follows.

Statistics for the deathpen_exonerations Variable	
Mean	
Median	
Mode	
Standard deviation	
Skewness	
Kurtosis	
Minimum	
Maximum	

   B. Create a histogram of deathpen_exonerations.[22] Override the default bar fill color with a color of your choice. Submit the histogram with your answers.

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

      Pundit 1's Pundit 2's

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

      _____

      _____

      _____

      _____

      _____

9. The States Workbook's dataset includes the percentage of women in state legislatures at different points in time. For this exercise, you'll compare the percentages of women in state legislatures in 2007 and 2017. These quantities are recorded by the womleg_2007 and womleg_2017 variables.

   A. Use the Histogram tool in Excel's Data Analysis ToolPak to create a frequency distribution table and histogram charts for the womleg_2007 and womleg_2017 variables.[23] To facilitate comparison, specify the histogram bins based on the following table. Based on the results, then complete the table.

	womleg_2007		womleg_2017	
Variable Values	Percentage	Frequency	Percentage	Frequency
0–5%		0		
5–10%		1		
10–15%		6		
15–20%		15		
20–25%		9		

[21]See Section 2.4.1 for guidance on generating tables of descriptive statistics with the Data Analysis ToolPak. In Section 1.2, we showed you how to activate Data Analysis add-ins.

[22]Section 2.4.2 shows you how to create histograms with Excel.

[23]See Section 2.4.2 for reference on creating histograms with the Data Analysis ToolPak. You'll need to use the histogram tool to customize bins (if you insert a histogram-type chart, you won't get this right).

Variable Values	womleg_2007		womleg_2017	
	Percentage	Frequency	Percentage	Frequency
25–30%		6		
30–35%		11		
35–40%		2		
40–45%		0		
45–50%		0		
Over 50%		0		
Total	100%	50	100.0%	50

B. Submit both of the histograms you created in part B. Label the histograms to indicate which one represents the variable values in 2007 and which one represents 2017.

10. 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. Apply the Descriptive Statistics tool in Excel's Data Analysis ToolPak to blackpct_2016 (the percentage of each state's population that is Black) and hispanicpct_2016 (the percentage of each state's population that is Hispanic).[24] Record the appropriate statistics for each variable in the table that follows.

	blackpct_2016	hispanicpct_2016
Mean		
Median		
Skewness		
Minimum		
Maximum		

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

Demographer 1 Demographer 2

Write a few sentences explaining your reasoning.

_____

_____

_____

_____

_____

_____

_____

C. Use the Data ▶ Sort procedure to obtain information on the percentage of Hispanics in the fifty states.[25]

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. _____

[24]Section 2.4.1 shows you how to generate tables of descriptive statistics with the Data Analysis ToolPak. Section 1.2 covers activating Data Analysis add-ins.

[25]See Section 2.5 for guidance on obtaining case-level information.