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

STATA COMMANDS AND FUNCTIONS USED

`codebook`	Provides detailed coding and labeling information about a dataset or variable[1]
`tabulate`	Generates a frequency distribution table (when applied to one variable)
`graph bar`	Creates bar charts of nominal or ordinal variables
`summarize`	Produces descriptive statistics about interval-level variables
`histogram`	Creates histograms of interval variables
`help`	Displays Stata manual information for a command
`sort`	Sorts the dataset by ascending values of a variable
`list`	Lists variables for cases in the dataset
`gsort`	Sorts the dataset by ascending or descending values of a variable

Descriptive statistics are the most basic—and sometimes the most informative—form of analysis you will do. Before you analyze why something varies, it's critical to understand how it is measured and how much it varies. In this chapter, you will learn how to use Stata to obtain descriptive statistics for variables in datasets.

Descriptive statistics are most often used to convey two attributes of a variable: its typical value (central tendency) and its spread (degree of dispersion or variation). You will find it helpful to describe variables using specialized terminology, tables of numbers, and graphics. Most empirical research begins with a description of the variables of interest.

Tutorials
Resources
Scan Code
2

[1] The underlined part of the command name shows you how it can be abbreviated in Stata. We write command names out completely in our examples for clarity, but you can abbreviate most commands to save time.

The precision with which we can describe central tendency for any given variable depends on the variable's **level of measurement**. For nominal-level variables, we can identify the *mode*, the most common value of the variable. For ordinal-level variables, those whose categories can be ranked, we 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, we can obtain the mode, median, and arithmetic *mean*, the sum of all values divided by the number of cases.

Finding a variable's central tendency is ordinarily a straightforward exercise. Simply read the computer results and report the numbers. Describing a variable's degree of dispersion or variation, however, often requires informed judgment.[2] 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. Using ordinary language, we might describe such a variable as "homogeneous." A variable has maximum dispersion if the cases are spread evenly across all values of the variable. The number of cases in one category equals the number of cases in every other category. In this circumstance, we would describe the variable as "heterogeneous."

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 economic system they believe to be the best: capitalism, communism, or socialism. What would be the modal response, the economic system preferred by most people? Capitalism. Would there be a great deal of dispersion, with large numbers of people choosing the alternatives, communism or socialism? Probably not.

If you ask many citizens a different question, 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 more evenly spread out 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, there are likely to be many people at either extreme: "very interested" and "not very interested." This would be an instance in which the amount of dispersion in a variable—its degree of spread—is essential to understanding and describing it.

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.

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 are three 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, 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 the people.

Some variables have qualitative values. For example, when we ask someone where they were born, their response is a place, like 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," that 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.

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

Some variables, like someone's age in years, provide quantitative information. Variables measured at the **interval level** provide precise, numerical information about the units of analysis. 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 units of analysis, 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. In the examples that follow, you will learn Stata's basic data description commands—`codebook`, `tabulate`, `summarize`, `sort`, and `list`.

The `tabulate` (abbreviated `tab`) and `summarize` (`sum`) commands are especially important and versatile. The `tabulate` command produces frequency distributions for nominal, ordinal, or interval variables. The `summarize` command returns descriptive statistics for interval-level variables. You will also learn to use `histogram` (`hist`), a command that creates graphic displays. By default, `histogram` produces descriptive graphs for interval variables. However, when supplied with the necessary options, `histogram` generates attractive bar charts for nominal and ordinal variables.

To generate basic data descriptions using commands like `codebook`, `tabulate`, and `summarize`, you may wish to type and run the commands from the Command Window instead of using the Do-file editor. However, nicely optioned bar charts and histograms, created with the `graph` and `histogram` commands, require a fair amount of typing. You will want to type and run these commands from a Do-file. (Do-files were covered in Chapter 1.)

2.2 DESCRIBING NOMINAL VARIABLES

In this section you will obtain descriptive statistics for a nominal-level variable in the World Dataset, region, which identifies the region where countries are located. First, we will use the `codebook` command, introduced in Chapter 1, to get a better idea of the information this variable encodes. Open the World Dataset. In the Command Window (or from the Do-file editor), type and run `codebook region`. (We used the `codebook` command and showed you how to access it via Stata's graphical user interface [GUI] in "Introduction: Getting Started with Stata.")

```
* use "World.dta"
codebook region
```

Stata responds:

```
region                                                      Region name
-----------------------------------------------------------------------

                 type: numeric (long)
                label: region_encode

                range: [1,8]                       units: 1
        unique values: 8                       missing .: 0/169

           tabulation:  Freq.  Numeric Label
                           48       1 Africa
                           27       2 Asia-pacific
                           28       3 C. Asia & E. Europe
                           19       4 Middle East
                            3       5 North America
                           24       6 South America
                            5       7 Scandinavia
                           15       8 Western Europe
```

Stata's response tells us that the region variable is encoded as numbers that range from 1 to 8. The region variable has eight unique numeric codes, each of which has a nonnumeric value label. Countries with region code 1, for example, are labeled "Africa," countries coded 5 are labeled "North America," and so on.

Some comments about how Stata stores data may be helpful at this point. Many Stata variables are stored with numeric values that are associated with text labels. For computational purposes, the region variable is encoded with numeric values ranging from 1 to 8. These values are associated with specific text labels (1=Africa, 2=Asia-Pacific, 3=C. Asia & E. Europe … 8=Western Europe). Unless instructed otherwise, Stata will use a variable's text labels (not its numeric values) when it displays results. If we wanted Stata to use region's numeric codes, we would enter this command: `tabulate region, nolabel` (we introduced the `tabulate` command in Chapter 1). The `nolabel` option suppresses value labels and instead displays the numeric codes.

Now we will ask Stata to produce a **frequency distribution table** for the region variable. (This kind of table is sometimes called a "one-way" frequency distribution). Type and run `tabulate region`. You can also execute the `tabulate` command with Stata's GUI by selecting Statistics ► Summaries, tables, and tests ► Frequency tables ► One-way table and then choosing region as the categorical variable.[3] The Results Window displays the frequency distribution of countries by region.

```
tabulate region

         Region name |      Freq.     Precent        Cum.
---------------------+-----------------------------------
              Africa |         48       28.40       28.40
        Asia-Pacific |         27       15.98       44.38
 C. Asia & E. Europe |         28       16.57       60.95
         Middle East |         19       11.24       72.19
       North America |          3        1.78       73.96
       South America |         24       14.20       88.17
         Scandinavia |          5        2.96       91.12
      Western Europe |         15        8.88      100.00
---------------------+-----------------------------------
               Total |        169      100.00
```

The value labels for each region code appear in the left-most column, with Africa occupying the top row of numbers and Western Europe the bottom row. There are three columns of numbers, labeled "Freq." (frequency), "Percent," and "Cum." (cumulative percent).

What does each column mean? The frequency column shows **raw frequencies**, the number of countries in each region. The percent column lists the percentage of countries in each category of the variable. So, for example, 48 of 169 total countries are located in Africa: 48/169 = 28.40%. Finally, the **cumulative percent** column reports the percentage of cases that fall in *or below* each value of the variable. For ordinal variables, as we will see, the cumulative percent column can provide valuable clues about how a variable is distributed. But for nominal variables, which cannot be ranked, the cumulative percent column provides no information of value.

Consider the percent column more closely. What is the **mode**, the most common region value? For nominal variables, the answer to this question is (almost) always an easy call: Simply find the value with the highest percentage of cases. Africa is the mode. There are more countries in Africa than there are in any other region. Does the region variable have a little dispersion or a lot of dispersion? Again, study the percent column. Apply the 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 countries located in Africa, or are there many countries in each region? Africa is the mode value, but it is not the region for the majority of countries. At the same time, the distribution

[3] We will typically only identify the GUI path to a Stata command when we first introduce a command, but we repeat it here for `tabulate` because it's such a workhorse command.

is not evenly spread out among regions; some region percentages are small (e.g., North America and Scandinavia). We would conclude that region has a medium level of dispersion.

2.3 DESCRIBING ORDINAL VARIABLES

Next, you will analyze and describe two ordinal-level variables in the NES Dataset, one of which has little variation and the other of which is more spread out. The American National Election Study (NES) is an important survey of Americans' political opinions, beliefs, and attitudes that's conducted every 2 years.

The NES Dataset contains the variable threat_from_china, which measures the extent to which Americans think the United States is threatened by China. Similarly, the variable threat_from_japan measures how much Americans think their country is threatened by Japan. Both variables have five possible values: not at all, a little, a moderate amount, a lot, and a great deal. Both variables are measured at the ordinal level.

Type and run two `tabulate` commands, one for threat_from_china (`tabulate threat _ from _ china [aw=wt]`) and one for threat_from_japan (`tabulate threat _ from _ japan [aw=wt]`). The `[aw=wt]` option tells Stata to weight observations using the NES Dataset's wt variable (see "A Closer Look: Weighted and Unweighted Analysis: What's the Difference?"). Stata will return two frequency distributions, one for each variable.

```
* use "NES.dta"
tabulate threat_from_china [aw= wt]
tabulate threat_from_japan [aw= wt]
```

Let's first focus on the frequency distribution for threat_from_china:

```
POST: How much is China a threat to the
                          United States |      Freq.     Percent        Cum.
----------------------------------------+-----------------------------------
                          1. Not at all |  415.053291       5.66        5.66
                           2. A little  | 848.1285625      11.56       17.22
                   3. A moderate amount |  2,094.7507      28.55       45.77
                               4. A lot |   1,737.463      23.68       69.45
                         5. A great deal|  2,241.6044      30.55      100.00
----------------------------------------+-----------------------------------
                                  Total |       7,337     100.00
```

How would you describe its central tendency and dispersion? Because threat_from_china is an ordinal variable, we can report both its mode and its **median**. Its mode is the response "5. A great deal," the option chosen by 30.55 percent of the sample. What about the median? This is where the cumulative percent column ("Cum.") of the frequency distribution comes into play. *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 at all") the median? No, this code contains fewer than half the cases. How about the second category ("2. A little")? No, again. According to the cumulative percent column, only 17.22 percent of the cases fall in or below this response category. It is not until we notch up in rank to the fourth category ("4. A lot") that the cumulative percentage exceeds the magic number of 50 percent. Because more than 50 percent of the cases fall in or below the fourth category (the cumulative percentage is equal to 69.45 percent), "4. A lot" is the median.

How would you describe the variation in Americans' perception of threat from China? If threat_from_china has a high level of variation, then the percentages of respondents in each category would be about equal. The modal value is "a great deal" of threat, but less than half of respondents (30.55 percent) gave this response. Many respondents said the threat level was "moderate" (28.55 percent) or "a lot" (23.68 percent). This variable has a fairly high level of dispersion.

Now examine Stata's frequency distribution table for threat_from_japan:

```
POST: How much is Japan a threat to the |
                          United States |      Freq.   Percent     Cum.
----------------------------------------+-----------------------------------
                          1. Not at all | 3,833.7688     52.42    52.42
                           2. A little  | 1,575.2586     21.54    73.96
                    3. A moderate amount| 1,350.2411     18.46    92.43
                               4. A lot | 336.921154      4.61    97.04
                         5. A great deal| 216.810403      2.96   100.00
----------------------------------------+-----------------------------------
                                  Total |      7,313    100.00
```

More than half of respondents (52.42 percent) fall into the modal value: "not at all." In contrast, few respondents said "a lot" (4.61 percent) or "a great deal" (2.96 percent). If threat_from_japan had no dispersion, then all the observations would fall into one category. That is, one value would have 100 percent of the cases, and each of the other categories would have 0 percent. This variable has a low level of dispersion, certainly less dispersion than the threat_from_china variable does.

Terms like *fairly high* or *moderate* are weak descriptions of dispersion, but measures of dispersion are limited when it comes to ordinal-level variables. You can describe the range of observed values as well as the interquartile range (IQR) of an ordinal-level variable (more on IQR in Section 2.5). As you'll see in the next section, with a higher level of measurement, we can describe a variable's dispersion more precisely.

A CLOSER LOOK

WEIGHTED AND UNWEIGHTED ANALYSIS: WHAT'S THE DIFFERENCE?

Many of this book's guided examples and exercises use the two survey datasets: the General Social Survey (GSS) and the American National Election Study (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 NES 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 Stata commands, including `recode` and `generate`, this lack of representativeness does not matter. For most Stata commands, however, the raw data produce incorrect results.

Fortunately, survey designers included the necessary corrective in the NES and GSS Datasets: a weight variable. **Weighted data** adjust for the distorting effect of sampling bias and help us calculate 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. When analyzing the GSS Dataset, you will specify the weight variable, wtss. For the NES Dataset, the weight variable is wt. Here and in Chapters 3–5, you will use the weight variables as *analytic weights*. In Chapters 6–11, you will use them as *probability weights*. The GSS and NES Datasets come from surveys and are meant to be analyzed with weighted observations. You don't always need to weight observations, however. You will often analyze **unweighted data**. When you're analyzing the Debate, States, or World Datasets, you don't need to use weights (and these datasets don't have variables for weighting observations).[4]

[4] These comments are generalizations. Some surveys and some analyses of surveys don't call for weights. Likewise, some analyses of aggregate-level data do call for weights (e.g., a population-weighted analysis of countries).

2.4 BAR CHARTS FOR NOMINAL AND ORDINAL VARIABLES

Thus far you have learned to wring a fair amount of information out of a dry handful of numbers: mode, median, and mean. Visual displays can add richness and nuance to these numerical descriptions of central tendency and variation. Two related types of graphs provide appropriate support for descriptive statistics.

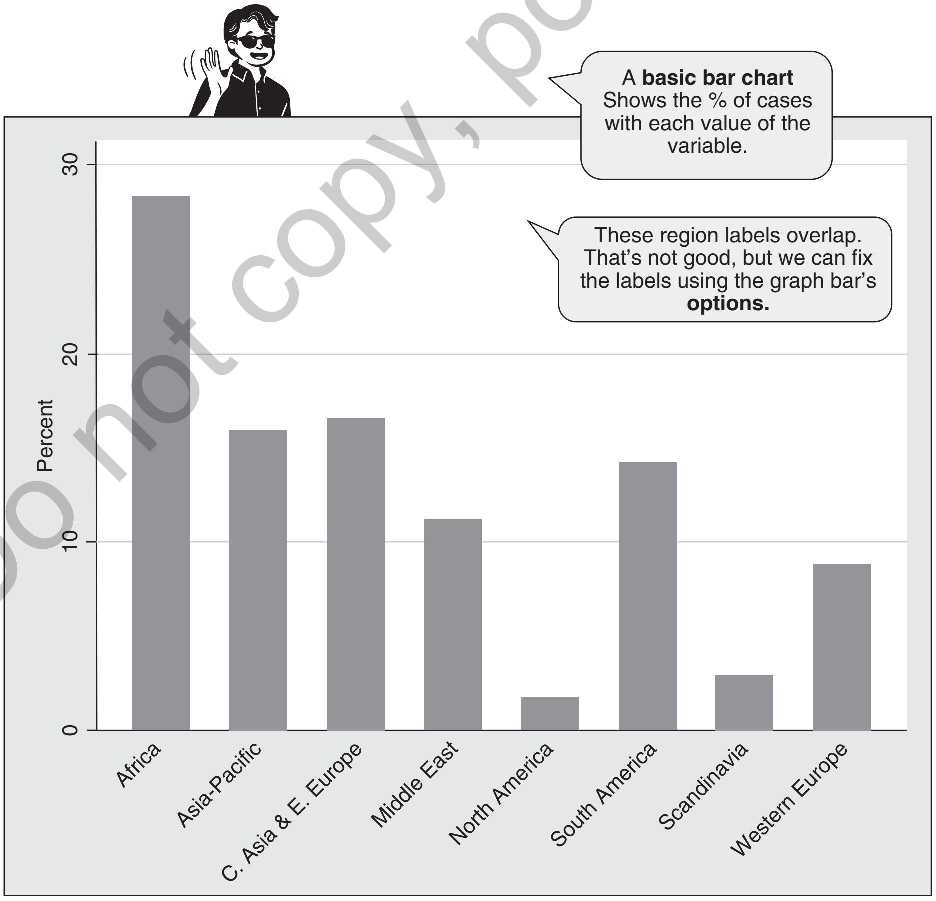
To visualize the central tendencies and distributions of nominal and ordinal variables, we can use bar charts. A **bar chart** displays each value of a variable and shows you the percentage of cases that fall into each category. Bar charts can also be used for interval variables that have a manageable number of values—generally 30 or fewer.

To create a bar chart for a variable, enter `graph bar, over(variable _ name)`, where `variable _ name` is the name of the variable to be graphed. You can also create a bar graph by selecting Graphics ► Bar chart. From the graph bar dialog's Main tab, choose Type of data: Graph of percent of frequencies within categories; from the Categories tab, select the variable to be graphed as Group 1's Grouping variable.

```
* use "World.dta"
graph bar, over(region)
```

The default bar chart rendering shown in Figure 2.1 is kind of a mess. There are eight tick marks along the x-axis, one for each category of region. The ticks are labeled using the region variable's value labels, from "Africa" (numeric code 1 on region) to "Western Europe" (code 8). Surely you noticed, however, that the variable labels—the names of the regions—overlap each other, making them difficult to read. The light-bluish gray background is also a bit of an eyesore.

FIGURE 2.1 ■ Bar Chart, Basic Version

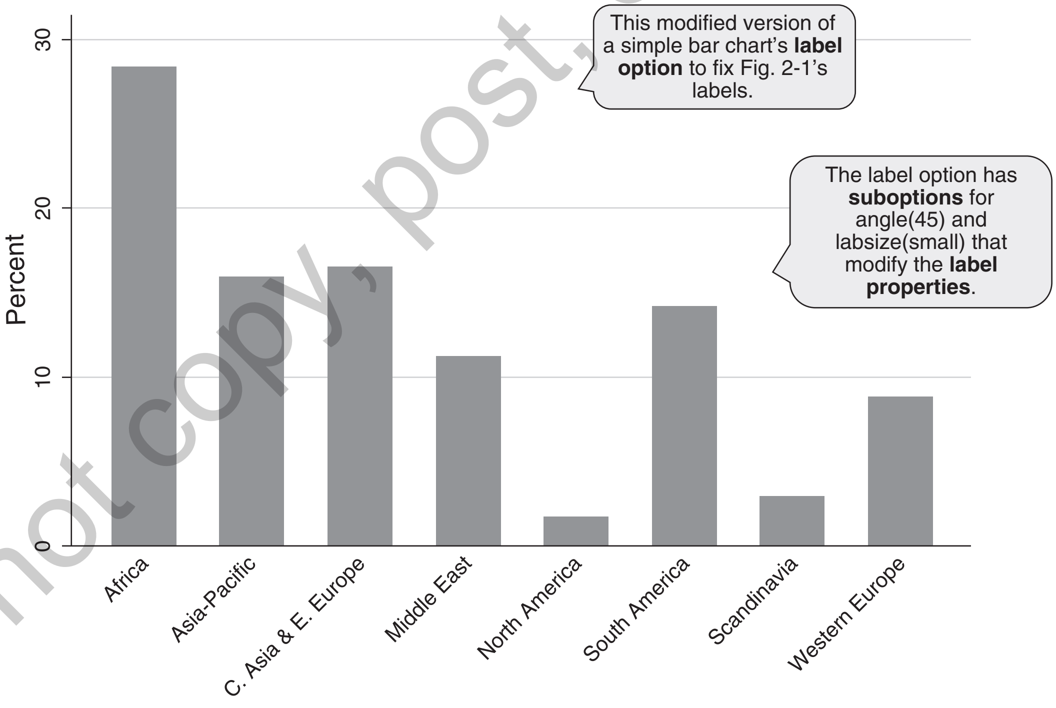


Fortunately, it is not too difficult to change bar chart settings to improve the visualization. When you want to customize graphics, we recommend you use the Do-file editor. Getting graphics right often requires some trial and error. You may want to experiment with different settings and may not want to retype commands (or start from scratch the next time). To improve the look of our region bar chart, let's rotate the variable labels 45 degrees and change the background color to white. (As you progress through the chapter and add commands to the Do-file, make sure to save it frequently.)

```
 Do-file Editor
#delimit ;
graph bar, over(region,
  label(angle(45) labsize(small)))
  graphregion(color(white))
;
```

The end result of our enhancements appears in Figure 2.2. As in our earlier effort, the y-axis records the percentage of countries falling into each category of region. Because each tick is labeled—and because some of the label names, such as "C. Asia & E. Europe" and "Western Europe," are pretty long—the value labels appear in smallish font, and they are angled at 45 degrees. These adjustments help to make the labels readable. Plus, they don't run into each other, as they do when displayed horizontally along the axis.

FIGURE 2.2 ■ Bar Chart, Enhanced with Options



This modified version of a simple bar chart's **label option** to fix Fig. 2-1's labels.

The label option has **suboptions** for angle(45) and labsize(small) that modify the **label properties**.

Each of these x-axis features—the labeling of each value, the use of word labels, the 45-degree angle, the small label font size—requires special suboptions within the bar chart's `over` option. Indeed, presentation-quality Stata graphs almost always require a fair number of (sometimes esoteric) options.

Take a few minutes to study the syntax for creating the improved bar chart. (For readability, the example uses the semicolon delimiter and adds extra spaces.) The first line includes the syntax of the basic chart: `graph bar, over (region)`. The `over(region)` option contains `label` refinements. The first `label` suboption, `angle(45)`, logically enough, rotates the variable labels 45 degrees counterclockwise from the x-axis. The second suboption, `labsize(small)`, downsizes the large type that Stata would otherwise use for value labels.[5] Don't forget to enclose the `over` option's suboptions in parentheses. Again, we set the graph's background color to white using `graphregion(color(white))` so it appears seamless on a page. Make sure that your finished product looks like the bar chart in Figure 2.2.

2.5 DESCRIBING INTERVAL VARIABLES

We 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 us the exact quantity of the characteristic being measured.*

Because interval variables have the most precision, they can be described more completely than can nominal or ordinal variables. For any interval-level variable, we can report its mode, median, and arithmetic average, or *mean*. In addition to these measures of central tendency, we can make more sophisticated judgments about variation. The most common measures of the dispersion of interval variables are variance and standard deviation.

To illustrate how we can use Stata to describe the central tendency and dispersion of an interval-level variable, we will analyze infant mortality rates, an important internal-level measure of health outcomes in countries around the world. To begin, we'll apply Stata's `summarize` command to the World Dataset's infant_mortality variable. This command generates descriptive statistics. With the World Dataset open, type and run `summarize infant _ mortality`. You can also access this command by selecting Statistics ► Summaries, tables, and tests ► Summary and descriptive statistics ► Summary statistics.

```
* use "World.dta"
summarize infant_mortality
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
infant_mor~y |        166    26.81687    24.39836        1.6      101.4
```

Stata tells us the number of countries or observations in the data ("Obs"), the mean infant mortality rate in these countries, the standard deviation of the distribution ("Std. Dev."), and the minimum and maximum observed values. By running `summarize` with default options, you can get a quick and concise profile of any interval variable in the dataset.

To obtain a more detailed description of a variable, append the `detail` option to the `summarize` command. Edit `summarize infant _ mortality` to read `summarize infant _ mortality, detail`. (*Hint:* If you are working from the Command Window, press the Page Up key, which returns `summarize infant _ mortality` to the command line, where it can be easily modified.)

```
summarize infant _ mortality, detail
```

The `detail` option, not surprisingly, instructs Stata to provide a fuller description of the variable. Consider the output:

```
      Number infants dying before age one per 1,000 live
                              births
-------------------------------------------------------------
      Percentiles      Smallest
 1%          1.6            1.6
 5%          2.7            1.6
10%          3.2            2.1       Obs                 166
25%            6            2.1       Sum of Wgt.         166

50%        18.45                      Mean           26.81687
                        Largest       Std. Dev.      24.39836
75%         41.9           90.1
90%         64.7           93.8       Variance       595.2801
95%         72.2           96.1       Skewness       1.004335
99%         96.1          101.4       Kurtosis       3.124455
```

[5] For a complete list of acceptable angle options, enter `help anglestyle`. For Stata's lengthy list of permissible label font sizes, enter `help textsizestyle`.

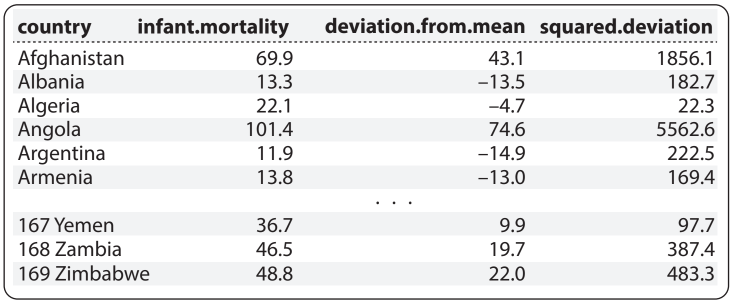
The mean infant mortality rate, 26.81687 (or 26.8), is again on display. Now notice the two left-hand columns of numbers, under the heading "Percentiles." Percentiles are synonymous with cumulative percentages. So, the left-most column displays cumulative percentages in ascending order (1%, 5%, 10% … 99%). The next column lists corresponding values of the infant mortality rate variable. Thus, the pairing "25% 6" can be read, "25 percent of countries have infant mortality rates of 6 or less." The pairing "50% 18.45" means, "50 percent of countries have infant mortality rates of 18.45 or less." The median is the same as the 50th percentile. To find the median of an interval-level variable, run `summarize` with the `detail` option and look for the value of the variable that is associated with "50%" in the percentiles column.

When you're analyzing an interval-level variable, like infant mortality rates, be skeptical about the reported mode(s). The values that appear to be modes are often an artificial by-product of rounding the quantity of interest to a small number of decimal places. Rounded to one decimal place, we have a few countries tied at 3.1, 3.2, and 3.3, but they do not really have the same infant mortality rates. If the variable's values are integers that count something, the mode might provide useful information; when you're analyzing a continuous variable with numbers after a decimal point, the mode usually doesn't help describe the variable's central tendency.

With variables measured at the interval level, we can measure dispersion more precisely than we can with nominal- or ordinal-level variables. One way to measure an interval-level variable's dispersion is its **range**. An interval-level variable's range is equal to the difference between its maximum and minimum observed values. The range of the World Dataset's infant mortality variable is 99.8, which is equal to the difference between its maximum value, 101.4, and its minimum value, 1.6.

Stata's `summarize` command with the `detail` option is so meticulous in providing percentiles that the numbers permit us to determine the **interquartile range**. The IQR comprises the values of a variable that bracket the "middle half" of a distribution, between the top of the lowest quartile ("25%") and the bottom of the highest quartile ("75%"). For the current example, we can see that the middle half of the distribution of infant mortality rates falls between 6 and 41.9 deaths per 1,000 live births. Now, the IQR has limited analytic value for describing a single variable. However, it is quite useful when comparing two or more distributions. (This is illustrated with box plots in Chapter 5.)

The two most common measures of the dispersion of interval variables are **variance** and **standard deviation**. These measures of dispersion express the typical amount of variation one observes in the values of a variable. Infant mortality rates around the world average 26.817, but the rates observed in individual countries are higher or lower than that. The infant mortality rates observed in different countries around the world deviate from the variable's mean value. The rate in countries like Afghanistan and Angola is higher than the mean; the rate in countries like Albania and Argentina is lower than the mean. In the following image, you can see the infant mortality rates observed in some countries, how much those observed values deviate from the variable's mean, and the result of squaring that deviation.



country	infant.mortality	deviation.from.mean	squared.deviation
Afghanistan	69.9	43.1	1856.1
Albania	13.3	–13.5	182.7
Algeria	22.1	–4.7	22.3
Angola	101.4	74.6	5562.6
Argentina	11.9	–14.9	222.5
Armenia	13.8	–13.0	169.4
		. . .	
167 Yemen	36.7	9.9	97.7
168 Zambia	46.5	19.7	387.4
169 Zimbabwe	48.8	22.0	483.3

Variance is the typical amount of squared deviation from the variable's mean. A function like `summarize` calculates variance by summing all the squared deviations and dividing that sum by the sample size – 1 (N – 1). Variance is not exactly the mean squared deviation but it's close to it, especially with large samples. The infant mortality variable's variance is 595.28.

Standard deviation is the square root of variance; it tells you how much absolute deviation from the mean is typical in the observations. Standard deviation is not exactly mean absolute deviation; it's a little bit larger than mean absolute deviation and a good estimate of typical deviation in the population (see Chapter 8 for a discussion of sample versus population). Both variance and standard deviation are reported in the table of descriptive statistics. The infant mortality variable's standard deviation is 24.398.

What is skewness and how do you know it when you see it? **Skewness** refers to how symmetrical a distribution is. 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 a small number of cases occupy extremely high values of an interval variable—distributions with a longer, skinnier right-hand tail—have a *positive skew*. If the distribution has a few cases at the extreme lower end—the distribution has a longer, skinnier left-hand tail—then the distribution has a *negative skew*. Infant mortality rates are positively skewed; the variable's skewness is 1.004.

When a distribution is highly skewed, it is a good practice to use the median instead of the mean in describing central tendency. Skewness has a predictable effect on the mean. A positive skew tends to pull the mean upward; a negative skew pulls it downward. However, skewness has less effect on the median. Since the median reports the middle-most value of a distribution, it is not tugged upward or downward by extreme values.

In addition, Stata's descriptive statistics table includes a statistic called **kurtosis**. Kurtosis measures whether the tails of a distribution are heavier (positive kurtosis values) or lighter (negative kurtosis values) than normal. However, we won't do much with kurtosis here, as this is outside the scope of this book.

A CLOSER LOOK

STATA'S GRAPH EDITOR

To modify graphics, Stata offers an alternative to specifying options and suboptions in a Do-file: the Graph Editor. To open the Graph Editor, click the Start Graph Editor icon on the toolbar above a Stata graphic, as shown in Figure 2.3. Take a few minutes to take stock of the editing environment.

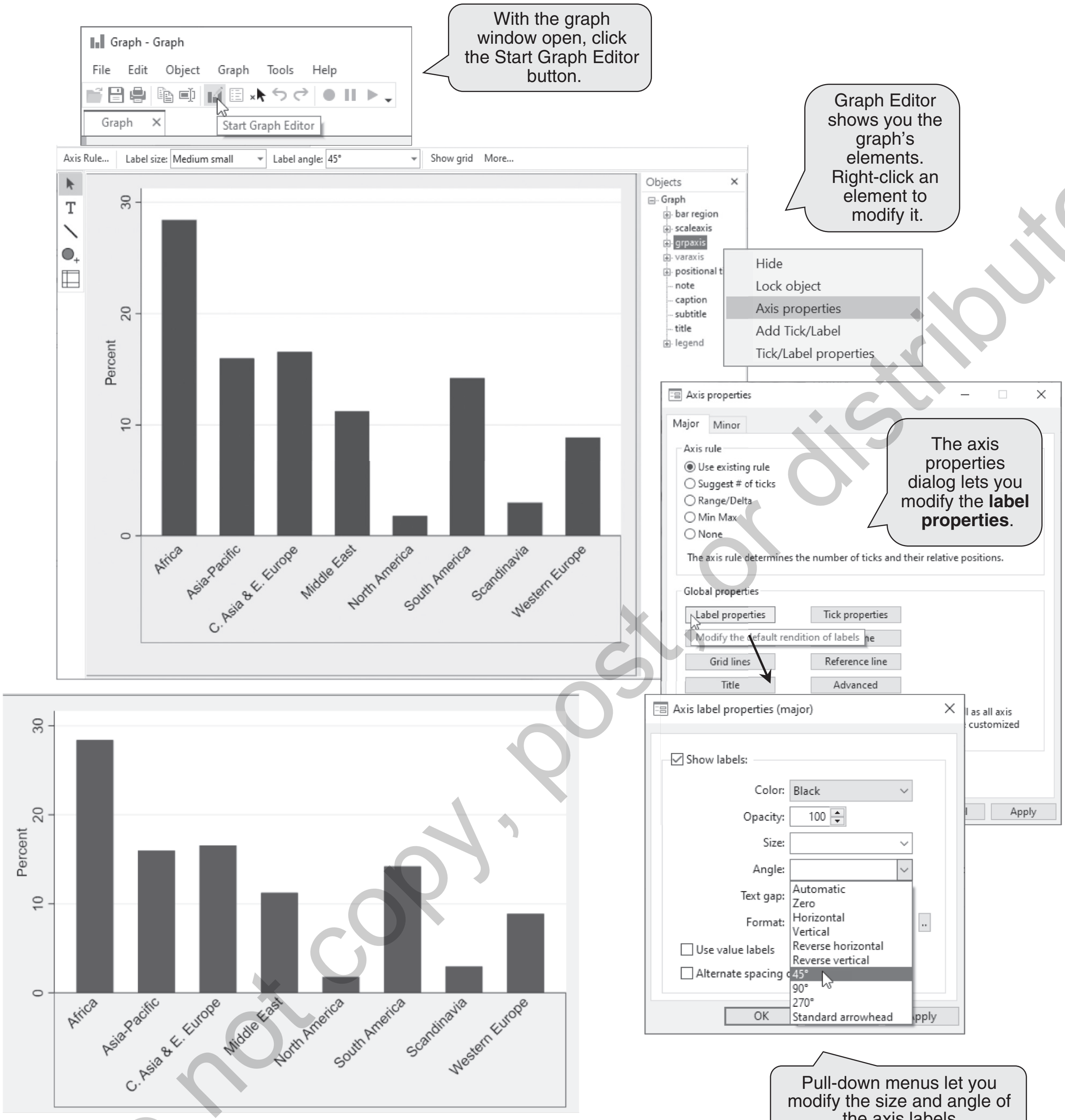
In a nutshell, here is how the Graph Editor works. The user selects an object for editing, which the editor outlines in red. The Contextual Toolbar, the main workhorse for most tasks, displays the editable features of the selected object. To make revisions, the user clicks drop-downs or types text on the Contextual Toolbar. For objects that already exist within the graph, you can select them with the mouse by simply clicking on them. The Object Browser, which occupies the right-hand margin of the editor, may be used to select existing objects or to add content such as text, titles, and notes. For example, to modify the x-axis labels in Figure 2.1 so they aren't overlapping, we first would bring the mouse to the Object Browser and click on "grpaxis," which is already highlighted in Figure 2.3. Stata selects the x-axis labels, outlining them in red. The Contextual Toolbar responds to our choice by showing us the Axis properties we can change quickly (Axis rule) and those that require some extra clicks, like the Label properties.

On the Contextual Toolbar, click Label properties to open the Label Properties Window (Figure 2.3). This window shows the `lab` options, including the size and angle for the x-axis labels. In this example, set the label size to small and angle to 45°. Click OK. Stata repositions the x-axis labels to make them more readable. To change the graph background to white, double-click "Graph" on the right-hand Object browser, change the color from Stata's default "Light-bluish gray" to "White," and then click OK. When you're satisfied with your edited graphic, click the Graph Editor button again the stop Editor and return to the graphics window.

The Graph Editor is a convenient tool for quickly editing graphs and a good way to learn about graphic elements and their options. The downside of the Graph Editor, however, for political analysis is that you cannot save graphics commands in a Do-file to use again or repurpose for other projects.[6] If you use the Graph Editor and later decide the graphic needs a subtitle or some other

[6] You can select Graphics ► Manage graphs ► Describe graph to obtain the commands used to create a graph stored in memory or saved to disk, but Stata won't show coding for options added using Graph Editor.

FIGURE 2.3 ■ Using Stata's Graph Editor



Label size: Medium small

Label angle: 45°

● Use existing rule

The axis rule determines the number of ticks and their relative positions.

☑ Show labels:

Color: Black

Opacity: 100

Size:

Angle:

Text gap:

Format:

☐ Use value labels

☐ Alternate spacing

With the graph window open, click the Start Graph Editor button.

Graph Editor shows you the graph's elements. Right-click an element to modify it.

The axis properties dialog lets you modify the **label properties**.

Pull-down menus let you modify the size and angle of the axis labels.

Stata's Graph Editor lets you point-and-click instead of writing graph commands.

change, you'll have to do all your edits over again. Unlike Stata's drop-down GUI for analysis commands, Stata's Graph Editor doesn't output the command-line equivalent of edited graphics to "reverse engineer" graphics. The Graph Editor Recorder somewhat alleviates this problem, but recordings can't be edited like Do-files.

The Graph Editor Recorder, as its name suggests, keeps a saveable and retrievable record of the edits you make to your graphs. Certainly, the recorder is a handy way to capture all your changes to a specific graph: Switch it on when you start the editor and switch it off (and save the recording) before you stop the editor. This feature also provides an efficient way to record generic preferences that we want to apply to any new graph. For example, suppose we like small font labels and prefer blue bar chart bars to Stata's default. First, of course, we would start the Graph Editor and click the red recording button. After we make any generic edits (font sizes, bar colors, or labeling conventions), we could stop recording and save the record, giving it a descriptive file name. New graphs will benefit from these efforts. We would then start the Graph Editor and click the green arrow. By retrieving the earlier recording—Stata will quite helpfully anticipate this step—we can automatically apply the earlier edits to the current project.

2.6 HISTOGRAMS FOR INTERVAL VARIABLES

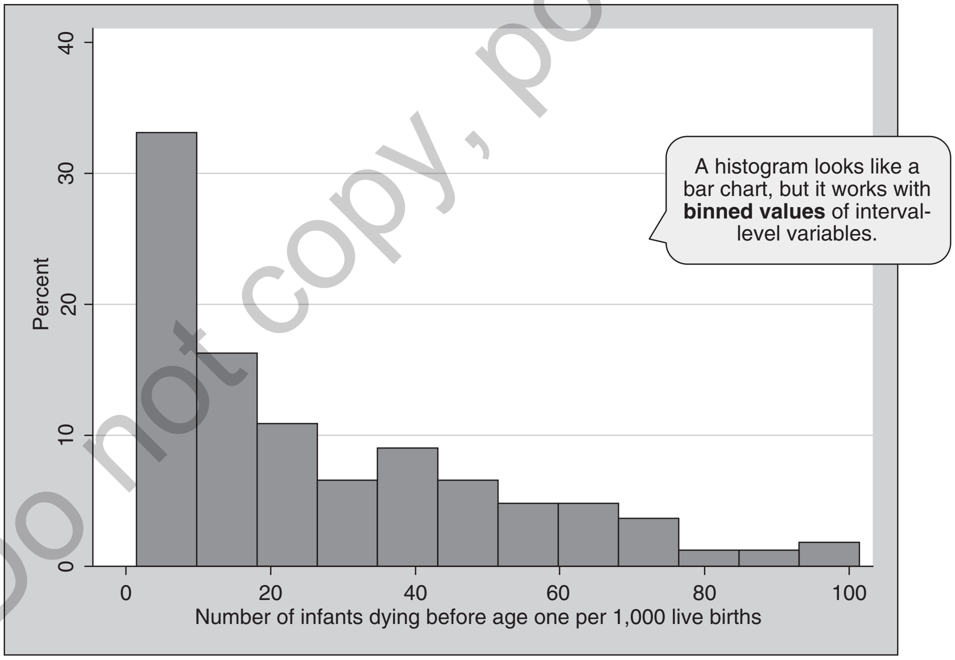
For an interval variable with many unique values, a histogram is generally a better choice than a bar chart. A **histogram** looks similar to a bar chart but instead of displaying each discrete value, it collapses categories into ranges called **bins**. Each bin covers a range of values, resulting in a compact display.

```
histogram varname, options
```

Now let's create a histogram of infant mortality rates observed in countries around the world—first a bare-sbones result, then a well-optioned graphic. To start, type and run `histogram infant _ mortality, percent`. You can also generate a histogram using Stata's GUI. Select Graphics ► Histogram and then choose the infant_mortality variable with the Y axis: Percent option. Stata delivers the simple histogram of infant mortality rates you see in Figure 2.4.

```
* use "World.dta"
histogram infant_mortality, percent
```

FIGURE 2.4 ■ Histogram, Basic Version



A histogram looks like a bar chart, but it works with **binned values** of interval-level variables.

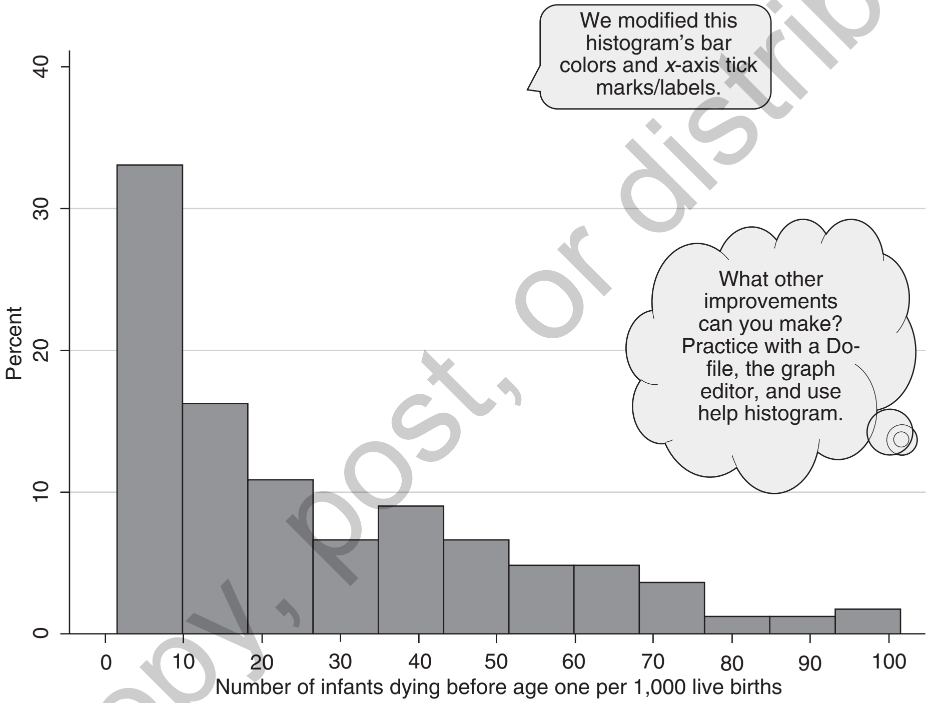
The histogram shows the % of countries in the 0–9 range, 9–18 range, and so on.

These bins come from default settings. You can change the bin ranges.

For quick, unadorned visual displays, these pithy commands work fine. Notice that the histogram's graphic signature—combining adjacent values of a variable—results in a compact, readable display. One can easily locate the center of the distribution (approximately 20). And the skinny right-hand tail reaffirms the presence of the positive skew we found in our numerical analysis. Stata automatically determines how many bins to include in the histogram if the number or width of bins is not specified. The axis titles ("Number of infants dying before age one per 1,000 live births" and "Percent") are acceptable, and the y-axis tick marks are nicely scaled and legible. For the product of a terse command running on Stata defaults, this graph is quite presentable as is.

A keen eye, however, might spot some issues with this histogram: The bar colors are bland, there are just a few x-axis ticks labeled, and the background color doesn't blend into the page. Let's open the Do-file editor and modify the histogram options to create a more useful graphic like Figure 2.5.

FIGURE 2.5 ■ Histogram, Enhanced with Options



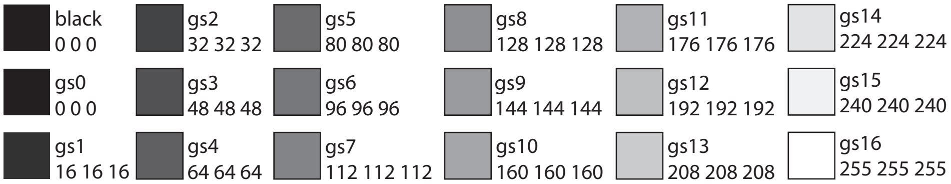
Suppose we prefer x-axis ticks labeled from 0 to 110 in increments of 10. Suppose further that we plan to print in color, and so we prefer a bolder fill color for the bars and want the histogram to appear seamless on a white page. Setting the color of the graph region to white with `graphregion(color(white))` makes a big difference visually. Finally, we'll use the `histogram` command's `xlabel` option to create more x-axis value labels.

Do-file Editor

```
#delimit ;
hist infant_mortality, percent
 fcolor(sand) color(sandb)
 graphregion(color(white))
 xlabel(0(10)100, labsize(medsmall) valuelabel)
 ;
```

The `histogram` command's `fcolor` and `color` options allow us to specify the fill and outline colors for the vertical bars (`sandb` stands for "bright sand" color). The settings for the `xlabel` option are somewhat tricky; the range-delta suboption, `0(10)100`, sets 0 as the lowest displayed value, 100 as the highest displayed value, and 10 as the increment. The x-axis will then read: "0 10 20 30 40 50 60 70 80 90 100." A nicely modified histogram appears above.

In the graphic we just created, the authors arbitrarily requested `sand` as the fill color. This may or may not be to your liking. To retrieve the names of all built-in colors, type `help colorstyle`. For many projects, grayscale colors, which range from `gs0` (black) to `gs16` (white), are particularly useful.[7]



There is one complication to note about making histograms: Stata does not permit analytic weights, the `[aw]` option, with the `histogram` command. However, it does permit frequency weights, the `[fw]` option. Frequency weights are the number of observations that a particular observation "stands for" in the sample; it is commonly a method of condensing a dataset with many duplicate observations. To accommodate Stata, we can round the sample weights to the nearest integer and use the rounded-off weights as frequency weights in the `histogram` command.[8] To do this, select Data ► Create or change data ► Create new variable. We cover variable transformations like this in more detail in Chapter 3. Make sure your histogram plots percentage frequencies on the y-axis (rather than counts).

2.7 OBTAINING CASE-LEVEL INFORMATION

When we analyze a large survey dataset, as we have just done, we generally are not interested in how respondent X or respondent Y answered a particular question. Rather, we want to know how the entire sample of respondents distributed themselves across the response categories of a variable. Sometimes, however, we gather data on particular cases because the cases are themselves inherently important. The States Dataset (50 cases) and World Dataset (169 cases) are good examples. With these datasets, we may want to push the descriptions beyond the relative anonymity of a `tabulate` analysis or a `summarize` command and find out where particular cases "are" on an interesting variable. Stata's `sort` and `list` commands are ready-made for such elemental insights.

Suppose that after we see the striking distribution of infant mortality rates around the world (see Section 2.5), we want to know more about the countries represented by the graphics and descriptive statistics. Which countries have the lowest infant mortality rates? Which have the highest rates?

By enlisting the `sort` command, we can sort countries on the basis of the infant_mortality variable's values, in descending order from highest to lowest (or in ascending order from lowest to highest).[9] The `list` command will then display the sorted values of infant_mortality along with, at our request, the name of each country. (The World Dataset contains the variable country, which records countries' names.)

With the World Dataset open, type and run the following command. You can also sort observations with Stata's `sort` command dialog, accessible by selecting Data ► Sort.

```
* use "World.dta"
sort infant_mortality
```

[7] In Stata, you can use RGB triplet coding to request custom colors. These numeric color codes are less memorable but offer virtually unlimited options. For example, the statement `fcolor("0 0 255")` produces the same color as `fcolor(blue)`. The statement `fcolor("93 97 255")` produces the attractive blue color featured in the SPSS color palette.

[8] Alternatively, for greater precision, one can create frequency weights by multiplying the weights variable by a number greater than 1 and then rounding the result to the nearest integer. This avoids the problem of weights less than 0.5 being rounded to 0. It is important to plot percentage frequencies, rather than counts, on the y-axis.

[9] A related Stata command, `gsort`, permits the user to sort observations in ascending or descending order. This command can be accessed by selecting Data ► Sort and choosing the "Advanced sort" option.

Stata silently sorts the countries from low infant mortality rates to high infant mortality rates.[10] To see the sorted infant mortality values along with the names of countries, type and run the following command. You can also select Data ► Describe data ► List data to execute this command using the `list` command dialog.

```
list infant _ mortality country
```

Stata responds by giving us a table that lists countries by ascending infant mortality rates.

```
     +----------------------------------+
     | infant~y                 country |
     |----------------------------------|
  1. |      1.6                 Iceland |
  2. |      1.6              Luxembourg |
  3. |      2.1                 Finland |
  4. |      2.1                   Japan |
  5. |      2.2               Singapore |
     |----------------------------------|

                  rows 6 - 160 omitted

       |-----------------------------------------------|
  161. |     78.3   Congo, Democratic Republic of the |
  162. |     88.4                                 Chad |
  163. |     90.1                              Somalia |
  164. |     93.8                         Sierra Leone |
  165. |     96.1             Central African Republic |
       |-----------------------------------------------|
  166. |    101.4                               Angola |
  167. |        .                             Maldives |
  168. |        .                               Taiwan |
  169. |        .                               Belize |
       +-----------------------------------------------+
```

We can see that Iceland, Luxembourg, Finland, Japan, and Singapore have the lowest infant mortality rates in the world. At the other end of the scale, Angola, Central African Republic, Sierra Leone, Somalia, and Chad have the highest infant mortality rates in the world.

The `sort` command is straightforward and intuitive, but it is limited to sorting observations in ascending order. If we want to list countries by descending values of infant_mortality so countries with the highest rates are listed first, the `gsort` command does the trick. Notice the minus sign, –, in front of infant_mortality in the following code; it specifies that we want the dataset sorted in descending order of infant_mortality. You can also select Data ► Sort (Advanced sort command) to execute this command using a dialog box.

```
gsort -infant_mortality
list infant_mortality country
```

```
     +-----------------------------------+
     | infant~y                  country |
     |-----------------------------------|
  1. |    101.4                   Angola |
  2. |     96.1 Central African Republic |
  3. |     93.8             Sierra Leone |
  4. |     90.1                  Somalia |
  5. |     88.4                     Chad |
     |-----------------------------------|
```

[10] To sort observations on additional criteria to break ties between observations with the same variable value, simply add other variables to the `sort` command.

```
                     rows 6 - 160 omitted

          |                                      |
    161.  |     2.3                     Slovenia |
    162.  |     2.2                    Singapore |
    163.  |     2.1                        Japan |
    164.  |     2.1                      Finland |
    165.  |     1.6                   Luxembourg |
          |--------------------------------------|
    166.  |     1.6                      Lceland |
    167.  |       .                     Maldives |
    168.  |       .                       Belize |
    169.  |       .                       Taiwan |
          +--------------------------------------+
```

You may notice that the three countries with missing infant mortality rate values (Maldives, Belize, and Taiwan) are listed at the end of the table. The infant mortality rates in these countries are neither high nor low; they are unknown and thus cannot be placed in ascending or descending order.

KEY TERMS

Bar chart
Bins
Cumulative percent
Frequency distribution table
Histogram
Interquartile range
Interval
Kurtosis
Level of measurement
Median

Mode
Nominal
Ordinal
Range
Raw frequencies
Skewness
Standard deviation
Unweighted data
Variance
Weighted data

CHAPTER 2 EXERCISES

1. How you analyze a variable depends on its level of measurement. To apply the right methods, you must be able to identify a variable's level of measurement.[11]

   A. The States Dataset includes a variable named min_wage, which reports 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 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

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

2. Practice identifying the level of measurement of variables by completing the following table.[12]

Dataset	Variable	What Does the Variable Measure?	Level of Measurement (select one)
States	voter_id_law		□ Nominal □ Ordinal □ Interval
States	opioid_rx_rate		□ Nominal □ Ordinal □ Interval
World	gender_equal3		□ Nominal □ Ordinal □ Interval
World	gender_inequality		□ Nominal □ Ordinal □ Interval
World	typerel		□ Nominal □ Ordinal □ Interval

3. 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; you can't add a column of cumulative percentages to the frequency distribution table of a nominal-level variable.[13] Why is that?

4. 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.[14] 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?

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

[13] 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.

[14] We discuss histograms as an alternative to bar charts in Section 2.6.

5. According to the Inter-Parliamentary Union, an international organization of parliaments, 23.7 percent of members of the U.S. House of Representatives are women.[15] How does the United States compare to other democratic countries? Is 23.7 percent comparatively low, comparatively high, or typical for a national legislature? The World Dataset contains a variable named womenleg, which records the percentage of women in the lower house of the legislature in each of 168 democracies.

A. Use Stata's `summarize` command to obtain descriptive statistics for womenleg. Using the command's output, fill in the blanks: The mean of womenleg is equal to ________. The median of womenleg is equal to __________. The minimum is equal to ________ and the maximum is equal to __________.

B. Analysts generally prefer to use the mean to summarize a variable's central tendency, except in cases where the mean gives a misleading indication of the true center of the distribution. Make a considered judgment. For womenleg, can the mean be used or should the median be used instead? (select one)

Mean Median

Explain your answer.

____

____

____

____

C. Recall that 23.7 percent of U.S. House members are women. 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 one-fourth of democracies in the percentage of female members." According to the output from the `summarize` command with `detail` option, women would need to constitute what percentage of the House to meet this goal? (select one)

About 15 percent About 30 percent About 45 percent

D. Use Stata's `sort` and `list` commands to obtain information on the percentages of women in legislatures around the world.[16]

Which five countries have the highest percentages of women in their legislatures?

1. ______________________________

2. ______________________________

3. ______________________________

4. ______________________________

5. ______________________________

Which five countries have the lowest percentages of women in their legislatures?

1. ______________________________

2. ______________________________

3. ______________________________

4. ______________________________

5. ______________________________

E. Create a nicely labeled histogram of the womenleg variable. Give the horizontal axis (x-axis) the following label: "Percentage of Women in Legislature." Give the chart this main title: "Percentage Women Legislators in 168 Democracies." Submit the histogram.

Additional exercises on fillable PDF forms are available to instructors online at edge.sagepub.com/pollock

[15] See the Inter-Parliamentary Union website at https://www.ipu.org.

[16] See Section 2.7 for guidance on obtaining case-level information. You may want to use the `gsort` command as well.