6 Getting to Know Your Data

OVERVIEW

Descriptive statistics and descriptive graphs are what they sound like – they are tools that describe variables. These tools are valuable because they can summarize a tremendous amount of information in a succinct fashion. In this chapter we discuss some of the most commonly used descriptive statistics and graphs, how we should interpret them, how we should use them, and their limitations.

6.1 GETTING TO KNOW YOUR DATA STATISTICALLY

Thus far we have discussed details of the measurement of variables. A lot of thought and effort goes into the measurement of individual variables. But once a researcher has collected data and become familiar and satisfied with how the variables were measured, it is important for them to get a good idea of the types of values that the individual variables take on before moving to testing for causal connections between two or more variables. For example, researchers might want to know, among other things: What do "typical" values for a variable look like? How tightly clustered (or widely dispersed) are these values?

Before proceeding to test for theorized relationships *between* two or more variables, it is essential to understand the properties and characteristics of each variable. To put it differently, we want to learn something about what the values of each variable "look like." How do we accomplish this? One possibility is to list all of the observed values of a measured variable. For example, the following are the percentages of popular votes for major party candidates that went to the candidate of the party of the sitting president during US presidential elections from 1876 to 2016:[1]

[1] This measure is constructed so that it is comparable across time. Although independent or third-party candidates have occasionally contested elections, we focus on only those votes

48.516, 50.220, 49.846, 50.414, 48.268, 47.760, 53.171, 60.006, 54.483, 54.708, 51.682, 36.148, 58.263, 58.756, 40.851, 62.226, 54.983, 53.778, 52.319, 44.710, 57.094, 49.913, 61.203, 49.425, 61.791, 48.951, 44.842, 59.123, 53.832, 46.379, 54.737, 50.262, 51.233, 46.311, 52.010, 51.111. We can see from this example that, once we get beyond a small number of observations, a listing of values becomes unwieldy. We will get lost in the trees and have no idea of the overall shape of the forest. For this reason, we turn to descriptive statistics and descriptive graphs, to take what would be a large amount of information and reduce it to bite-size chunks that summarize that information.

YOUR TURN: Identifying a typical value

Based on the listing of values for incumbent vote, what do you think is a typical value for this variable?

Descriptive statistics and graphs are useful tools for helping researchers to get to know their data before they move to testing causal hypotheses. They are also sometimes helpful when writing about one's research. You have to make the decision of whether or not to present descriptive statistics and/or graphs in the body of a paper on a case-by-case basis. It is scientifically important, however, that this information be made available to consumers of your research in some way.[2]

One major way to distinguish among variables is the **measurement metric**. A variable's measurement metric is the type of values that the variable takes on, and we discuss this in detail in the next section by describing three different variable types. We then explain that, despite the imperfect nature of the distinctions among these three variable types, we are forced to choose between two broad classifications of variables – categorical or continuous – when we describe them. The rest of this chapter discusses strategies for describing **categorical variables** and **continuous variables**.

6.2 WHAT IS THE VARIABLE'S MEASUREMENT METRIC?

There are no hard-and-fast rules for describing variables, but a major initial juncture that we encounter involves the metric in which we measure each

for the two major parties. Also, because we want to test the theory of economic voting, we need to have a measure of support for incumbents. In elections in which the sitting president is not running for reelection, there is still reason to expect that their party will be held accountable for economic performances.

[2] Many researchers will present this information in an appendix (often made available online) unless there is something particularly noteworthy about the characteristics of one or more of their variables.

variable. Remember from Chapter 1 that we can think of each variable in terms of its label and its values. The label is the description of the variable – such as "Gender of survey respondent" – and its values are the denominations in which the variable occurs – such as "Male" or "Female." For treatment in most statistical analyses, we are forced to divide our variables into two types according to the metric in which the values of the variable occur: categorical or continuous. In reality, variables come in at least three different metric types, and there are a lot of variables that do not neatly fit into just one of these classifications. To help you to better understand each of these variable types, we will go through each with an example. All of the examples that we are using in these initial descriptions come from survey research, but the same basic principles of measurement metric hold regardless of the type of data being analyzed.

6.2.1 Categorical Variables

Categorical variables are variables for which cases have values that are either different from or the same as the values for other cases, but about which we cannot make any universally holding ranking distinctions. If we consider a variable that we might label "Religious Identification," some values for this variable are "Catholic," "Muslim," "nonreligious," and so on. Although these values are clearly different from each other, we cannot make universally holding ranking distinctions across them. More casually, with categorical variables like this one, it is not possible to rank order the categories from least to greatest: The value "Muslim" is neither greater nor less than "nonreligious" (and so on), for example. Instead, we are left knowing that cases with the same value for this variable are the same, whereas those cases with different values are different. The term "categorical" expresses the essence of this variable type; we can put individual cases into categories based on their values, but we cannot go any further in terms of ranking or otherwise ordering these values.

6.2.2 Ordinal Variables

Like categorical variables, **ordinal variables** are also variables for which cases have values that are either different from or the same as the values for other cases. The distinction between ordinal and categorical variables is that we *can* make universally holding ranking distinctions across the variable values for ordinal variables. For instance, consider the variable labeled "Retrospective Family Financial Situation" that has commonly been used as an independent variable in individual-level economic voting studies. In the 2004 National Election Study (NES), researchers created this variable

by first asking respondents to answer the following question: "We are interested in how people are getting along financially these days. Would you say that you (and your family living here) are better off or worse off than you were a year ago?" Researchers then asked respondents who answered "Better" or "Worse": "Much [better/worse] or somewhat [better/worse]?" The resulting variable was then coded as follows:

1. much better
2. somewhat better
3. same
4. somewhat worse
5. much worse

This variable is pretty clearly an ordinal variable because as we go from the top to the bottom of the list we are moving from better to worse evaluations of how individuals (and their families with whom they live) have been faring financially in the past year.

As another example, consider the variable labeled "Party Identification." In the 2004 NES researchers created this variable by using each respondent's answer to the question, "Generally speaking, do you usually think of yourself as a Republican, a Democrat, an Independent, or what?"[3] which we can code as taking on the following values:

1. Republican
2. Independent
3. Democrat

If all cases that take on the value "Independent" represent individuals whose views lie somewhere between "Republican" and "Democrat," we can call "Party Identification" an ordinal variable. If this is not the case, then this variable is a categorical variable.

YOUR TURN: Is that variable categorical or ordinal?

Choose a variable from the United States National Election Study 2016 post-election questionnaire (located at http://www.electionstudies.org/studypages/anes_timeseries_2016/anes_timeseries_2016_qnaire_post.pdf). Is that variable categorical or ordinal? Why?

[3] Almost all US respondents put themselves into one of the first three categories. For instance, in 2004, 1128 of the 1212 respondents (93.1 percent) to the post-election NES responded that they were a Republican, Democrat, or an Independent. For our purposes, we will ignore the "or what" cases. Note that researchers usually present partisan identification across seven values ranging from "Strong Republican" to "Strong Democrat" based on follow-up questions that ask respondents to further characterize their positions.

6.2.3 Continuous Variables

An important characteristic that ordinal variables *do not* have is **equal unit differences**. A variable has equal unit differences if a one-unit increase in the value of that variable *always* means the same thing. If we return to the examples from the previous section, we can rank order the five categories of "Retrospective Family Financial Situation" from 1 for the best situation to 5 for the worst situation. But we may not feel very confident working with these assigned values the way that we typically work with numbers. In other words, can we say that the difference between "somewhat worse" and "same" (4 – 3) is the same as the difference between "much worse" and "somewhat worse" (5 – 4)? What about saying that the difference between "much worse" and "same" (5 – 3) is twice the difference between "somewhat better" and "much better" (2 – 1)? If the answer to both questions is "yes," then "Retrospective Family Financial Situation" is a continuous variable.

If we ask the same questions about "Party Identification," we should be somewhat skeptical. We can rank order the three categories of "Party Identification," but we cannot with great confidence assign "Republican" a value of 1, "Independent" a value of 2, and "Democrat" a value of 3 and work with these values in the way that we typically work with numbers. We cannot say that the difference between an "Independent" and a "Republican" (2 – 1) is the same as the difference between a "Democrat" and an "Independent" (3 – 2) – despite the fact that both 3 – 2 and 2 – 1 = 1. Certainly, we cannot say that the difference between a "Democrat" and a "Republican" (3 – 1) is twice the difference between an "Independent" and a "Republican" (2 – 1) – despite the fact that 2 is twice as big as 1.

The metric in which we measure a variable has equal unit differences if a one-unit increase in the value of that variable indicates the same amount of change across *all values* of that variable. Continuous variables are variables that *do* have equal unit differences.[4] Imagine, for instance, a variable labeled "Age in Years." A one-unit increase in this variable *always* indicates an individual who is one year older; this is true when we are talking about a case with a value of 21 just as it is when we are talking about a case with a value of 55.

[4] We sometimes call these variables "interval variables." A further distinction you will encounter with continuous variables is whether they have a substantively meaningful zero point. We usually describe variables that have this characteristic as "ratio" variables.

YOUR TURN: Is the Polity IV measure of democracy continuous?

In the previous chapter, we discussed the way in which the Polity IV measure of democracy is constructed. Would you feel comfortable treating this measure as a continuous variable? Why or why not?

6.2.4 Variable Types and Statistical Analyses

As we saw in the preceding sections, variables do not always neatly fit into the three categories. When we move to the vast majority of statistical analyses, we must decide between treating each of our variables as though it is categorical or as though it is continuous. For some variables, this is a very straightforward choice. However, for others, this is a very difficult choice. If we treat an ordinal variable as though it is categorical, we are acting as though we know less about the values of this variable than we really know. On the other hand, treating an ordinal variable as though it is a continuous variable means that we are assuming that it has equal unit differences. Either way, it is critical that we be aware of our decisions. We can always repeat our analyses under a different assumption and see how robust our conclusions are to our choices.

With all of this in mind, we present separate discussions of the process of describing a variable's **variation** for categorical and continuous variables. A variable's variation is the distribution of values that it takes across the cases for which it is measured. It is important that we have a strong knowledge of the variation in each of our variables before we can translate our theory into hypotheses, assess whether there is covariation between two variables (causal hurdle 3 from Chapter 3), and think about whether or not there might exist a third variable that makes any observed covariation between our independent and dependent variables spurious (hurdle 4). As we just outlined, descriptive statistics and graphs are useful summaries of the variation for individual variables. Another way in which we describe distributions of variables is through measures of **central tendency**. Measures of central tendency tell us about typical values for a particular variable at the center of its distribution.

6.3 DESCRIBING CATEGORICAL VARIABLES

With categorical variables, we want to understand the frequency with which each value of the variable occurs in our data. The simplest way of seeing this is to produce a frequency table in which the values of the categorical variable are displayed down one column and the frequency

with which it occurs (in absolute number of cases and/or in percentage terms) is displayed in (an)other column(s). Table 6.1 shows such a table for the variable "Religious Identification" from the NES survey measured during the 2004 national elections in the United States.

The only measure of central tendency that is appropriate for a categorical variable is the **mode**, which is defined as the most frequently occurring value. In Table 6.1, the mode of the distribution is "Protestant," because there are more Protestants than there are members of any other single category.

Table 6.1 "Religious Identification" from the NES survey measured during the 2004 national elections in the United States

Category	Number of cases	Percent
Protestant	672	56.14
Catholic	292	24.39
Jewish	35	2.92
Other	17	1.42
None	181	15.12
Total	1197	99.99

A typical way in which non-statisticians present frequency data is in a pie graph such as Figure 6.1. Pie graphs are one way for visualizing the percentage of cases that fall into particular categories. Many statisticians argue strongly against their use and, instead, advocate the use of bar graphs. Bar graphs, such as Figure 6.2, are another graphical way to illustrate frequencies of categorical variables. It is worth noting, however, that most of the

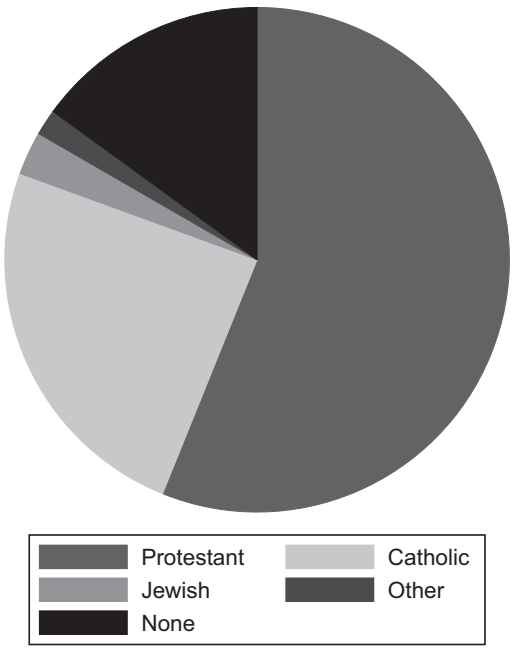

Figure 6.1 Pie graph of religious identification, NES 2004

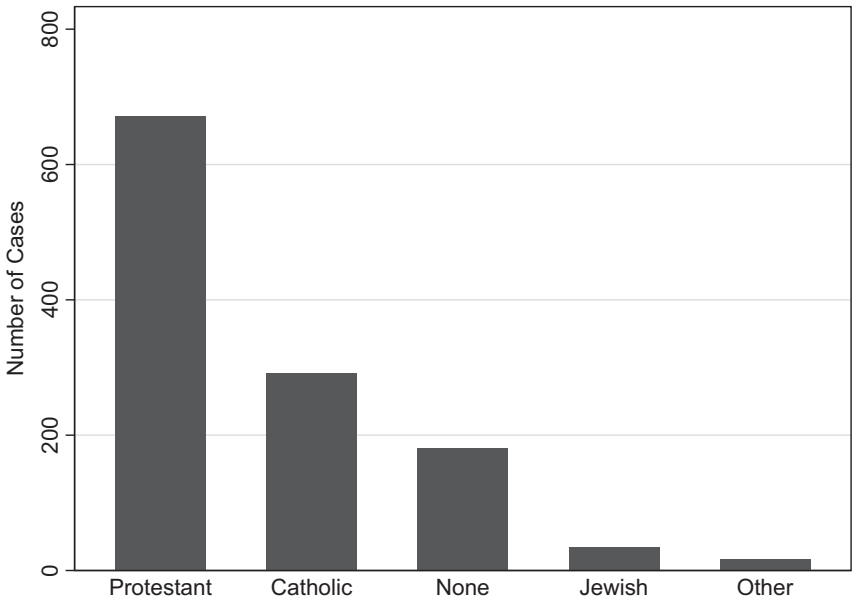

Figure 6.2 Bar graph of religious identification, NES 2004

information that we are able to gather from these two figures is very clearly and precisely presented in the columns of frequencies and percentages displayed in Table 6.1.

6.4 DESCRIBING CONTINUOUS VARIABLES

The statistics and graphs for describing continuous variables are considerably more complicated than those for categorical variables. This is because continuous variables are more mathematically complex than categorical variables. With continuous variables, we want to know about the central tendency and the spread or variation of the values around the central tendency. With continuous variables we also want to be on the lookout for **outliers**. Outliers are cases for which the value of the variable is extremely high or low relative to the rest of the values for that variable. When we encounter an outlier, we want to make sure that such a case is real and not created by some kind of error.

Most statistical software programs have a command for getting a battery of descriptive statistics on continuous variables. Figure 6.3 shows the output from Stata's "summarize" command with the "detail" option for the percentage of the major party vote won by the incumbent party in every US presidential election between 1876 and 2016.[5] The statistics

[5] These data come from a famous US forecasting model, developed by Ray Fair; see https://fairmodel.econ.yale.edu.

```
. summarize inc_vote, det

                        inc_vote
-------------------------------------------------------------
      Percentiles      Smallest
 1%       36.148          36.148
 5%       40.851          40.851
10%       44.842           44.71       Obs                  36
25%      48.7335          44.842       Sum of Wgt.          36

50%      51.4575                       Mean           51.92569
                        Largest       Std. Dev.      5.785544
75%        54.86          60.006
90%       60.006          61.203       Variance       33.47252
95%       61.791          61.791       Skewness      -.3039279
99%       62.226          62.226       Kurtosis       3.274385
```

Figure 6.3 Example output from Stata's "summarize" command with "detail" option

on the left-hand side (the first three columns on the left) of the computer printout are what we call **rank statistics**, and the statistics on the right-hand side (the two columns on the right-hand side) are known as the **statistical moments**. Although both rank statistics and statistical moments are intended to describe the variation of continuous variables, they do so in slightly different ways and are thus quite useful together for getting a complete picture of the variation for a single variable.

6.4.1 Rank Statistics

The calculation of rank statistics begins with the ranking of the values of a continuous variable from smallest to largest, followed by the identification of crucial junctures along the way. Once we have our cases ranked, the midpoint as we count through our cases is known as the median case. Remember that earlier in the chapter we defined the variable in Figure 6.3 as the percentage of popular votes for major party candidates that went to the candidate from the party of the sitting president during US presidential elections from 1876 to 2016. We will call this variable "Incumbent Vote" for short. To calculate rank statistics for this variable, we need to first put the cases in order from the smallest to the largest observed value. This ordering is shown in Table 6.2. With rank statistics we measure the central tendency as the **median value** of the variable. The median value is the value of the case that sits at the exact center of our cases when we rank them from the smallest to the largest observed values. When we have an even number of cases, as we do in Table 6.2, we average the value of the two centermost ranked cases to obtain the median value (in our example we calculate the median as $\frac{1}{2}(51.233 + 51.682) = 51.4575$). This is also known as the value of the variable at the 50 percent rank. In a similar way, we can talk about the value of the variable at any other

Table 6.2 Values of "Incumbent Vote" ranked from smallest to largest

Rank	Year	Value
1	1920	36.148
2	1932	40.851
3	1952	44.710
4	1980	44.842
5	2008	46.311
6	1992	46.379
7	1896	47.760
8	1892	48.268
9	1876	48.516
10	1976	48.951
11	1968	49.425
12	1884	49.846
13	1960	49.913
14	1880	50.220
15	2000	50.262
16	1888	50.414
17	2016	51.111
18	2004	51.233
19	1916	51.682
20	2012	52.010
21	1948	52.319
22	1900	53.171
23	1944	53.778
24	1988	53.832
25	1908	54.483
26	1912	54.708
27	1996	54.737
28	1940	54.983
29	1956	57.094
30	1924	58.263
31	1928	58.756
32	1984	59.123
33	1904	60.006
34	1964	61.203
35	1972	61.791
36	1936	62.226

percentage rank in which we have an interest. Other ranks that are often of interest are the 25 percent and 75 percent ranks, which are also known as the first and third "quartile ranks" for a distribution. The difference between the variable value at the 25 percent and the 75 percent ranks is known as the "interquartile range" or "IQR" of the variable. In our example variable, the 25 percent value is $\frac{1}{2}(48.516 + 48.951) = 48.7335$ and the 75 percent value is $\frac{1}{2}(54.737 + 54.983) = 54.8600$. This makes the IQR $= 54.8600 - 48.7335 = 6.1265$. In the language of rank statistics, the median value for a variable is a measure of its central tendency, whereas the IQR is a measure of the **dispersion**, or spread, of values.

With rank statistics, we also want to look at the smallest and largest values to identify outliers. Remember that we defined outliers at the beginning of this section as "cases for which the value of the variable is extremely high or low relative to the rest of the values for that variable." If we look at the highest values in Table 6.2, we can see that there aren't really any cases that fit this description. Although there are certainly some values that are a lot higher than the median value and the 75 percent value, they aren't "extremely" higher than the rest of the values. Instead, there seems to be a fairly even progression from the 75 percent value up to the highest value. The story at the lower end of the range of values in Table 6.2 is a little different. We can see that the two lowest values are pretty far from each other and from the rest of the low values. The value

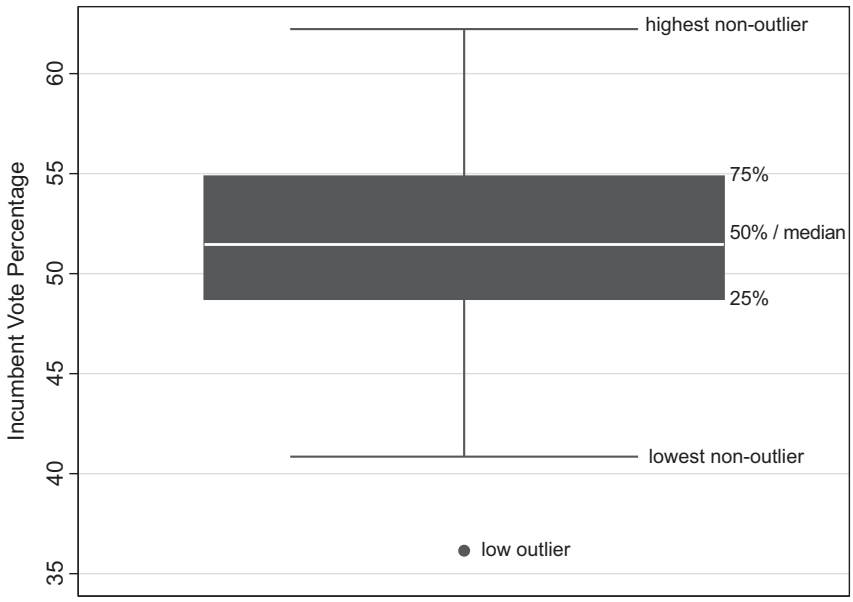

Figure 6.4 Box–whisker plot of incumbent-party presidential vote percentage, 1876–2016

of 36.148 in 1920 seems to meet our definition of an outlier. The value of 40.851 in 1932 is also a borderline case. Whenever we see outliers, we should begin by checking whether we have measured the values for these cases accurately. Sometimes we find that outliers are the result of errors when entering data. In this case, a check of our data set reveals that the outlier case occurred in 1920 when the incumbent-party candidate received only 36.148 percent of the votes cast for the two major parties. A further check of our data indicates that this was indeed a correct measure of this variable for 1920.[6]

Figure 6.4 presents a box–whisker plot of the rank statistics for our presidential vote variable. This plot displays the distribution of the variable along the vertical dimension. If we start at the center of the box in Figure 6.4, we see the median value (or 50 percent rank value) of our variable represented as the slight gap in the center of the box. The other two ends of the box show the values of the 25 percent rank and the 75 percent rank of our variable. The ends of the whiskers show the lowest and highest non-outlier values of our variable. Each statistical program has its own rules for dealing with outliers, so it is important to know whether

[6] An obvious question is "Why was 1920 such a low value?" This was the first presidential election in the aftermath of World War I, during a period when there was a lot of economic and political turmoil. The election in 1932 was at the very beginning of the large economic downturn known as "the Great Depression," so it makes sense that the party of the incumbent president would not have done very well during this election.

your box–whisker plot is or is not set up to display outliers. These settings are usually adjustable within the statistical program. The calculation of whether an individual case is or is not an outlier in this box–whisker plot is fairly standard. This calculation starts with the IQR for the variable. Any case is defined as an outlier if its value is either 1.5 times the IQR higher than the 75 percent value or if its value is 1.5 times the IQR lower than the 25 percent value. For Figure 6.4 we have set things up so that the plot displays the outliers, and we can see one such value at the bottom of our figure. As we already know from Table 6.2, this is the value of 36.148 from the 1920 election.

6.4.2 Moments

The statistical moments of a variable are a set of statistics that describe the central tendency for a single variable and the distribution of values around it. The most familiar of these statistics is known as the **mean value** or "average" value for the variable. For a variable Y, the mean value is calculated as

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n},$$

where $\bar{Y}$, read aloud as "Y-bar," indicates the mean of Y, which is equal to the sum of all values of Y across individual cases of Y, Y_i, divided by the total number of cases, n.[7] Although everyone is familiar with mean or average values, not everyone is familiar with the two characteristics of the mean value that make it particularly attractive to people who use statistics. The first is known as the "**zero-sum property**":

$$\sum_{i=1}^{n} (Y_i - \bar{Y}) = 0,$$

which means the sum of the difference between each Y value, Y_i, and the mean value of Y, $\bar{Y}$, is equal to zero. The second desirable characteristic of the mean value is known as the "**least-squares property**":

$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 < \sum_{i=1}^{n} (Y_i - c)^2 \quad \forall\, c \neq \bar{Y},$$

which means that the sum of the squared differences between each Y value, Y_i, and the mean value of Y, $\bar{Y}$, is less than the sum of the squared differences between each Y value, Y_i, and some value c, for all ($\forall$) c values not equal to ($\neq$) $\bar{Y}$. Because of these two properties, the mean value is

[7] To understand formulae like this, it is helpful to read through each of the pieces of the formula and translate them into words, as we have done here.

also referred to as the **expected value** of a variable. Think of it this way: If someone were to ask you to guess what the value for an individual case is without giving you any more information than the mean value, based on these two properties of the mean, the mean value would be the best guess.

The next statistical moment for a variable is the **variance** (var). We calculate the variance as follows:

$$\text{var}(Y) = \text{var}_Y = s_Y^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1},$$

which means that the variance of Y is equal to the sum of the squared differences between each Y value, Y_i, and its mean divided by the number of cases minus one.[8] If we look through this formula, what would happen if we had no variation on Y at all ($Y_i = \bar{Y}\ \forall\ i$)? In this case, variance would be equal to zero. But as individual cases are spread further and further from the mean, this calculation would increase. This is the logic of variance: It conveys the spread of the data around the mean. A more intuitive measure of variance is the **standard deviation** (sd):

$$\text{sd}(Y) = \text{sd}_Y = s_Y = \sqrt{\text{var}(Y)} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}}.$$

Roughly speaking, this is the average difference between values of Y (Y_i) and the mean of Y ($\bar{Y}$). At first glance, this may not be apparent. But the important thing to understand about this formula is that the purpose of squaring each difference from the mean and then taking the square root of the resulting sum of squared deviations is to keep the negative and positive deviations from canceling each other out.[9]

The variance and the standard deviation give us a numerical summary of the distribution of cases around the mean value for a variable.[10] We can also visually depict distributions. The idea of visually depicting distributions is to produce a two-dimensional figure in which the horizontal

[8] The "minus one" in this equation is an adjustment that is made to account for the number of "degrees of freedom" with which this calculation was made. We will discuss degrees of freedom in Chapter 8.

[9] An alternative method that would produce a very similar calculation would be to calculate the average value of the absolute value of each difference from the mean: $\left(\frac{\sum_{i=1}^{n}|Y_i - \bar{Y}|}{n}\right)$.

[10] The **skewness** and the **kurtosis** of a variable convey the further aspects of the distribution of a variable. The skewness calculation indicates the symmetry of the distribution around the mean. If the data are symmetrically distributed around the mean, then this statistic will equal zero. If skewness is negative, this indicates that there are more values below the mean than there are above; if skewness is positive, this indicates that there are more values above the mean than there are below. The kurtosis indicates the steepness of the statistical distribution. Positive kurtosis values indicate very steep distributions, or a concentration of values close to the mean value, whereas negative kurtosis values indicate a flatter distribution, or more cases further from the mean value. For the normal distribution, which we will discuss in Chapter 7, skewness = 0 and kurtosis = 3.

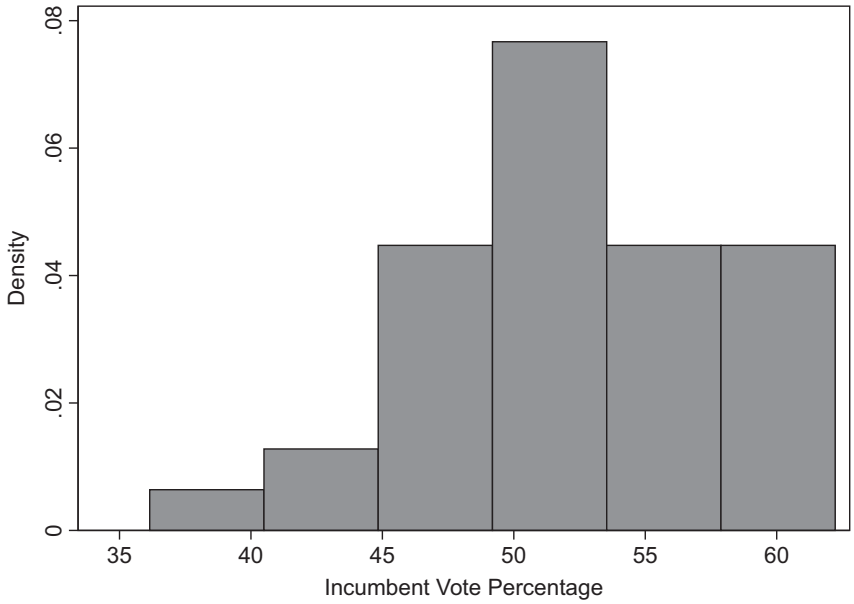

Figure 6.5 Histogram of incumbent-party presidential vote percentage, 1876–2016

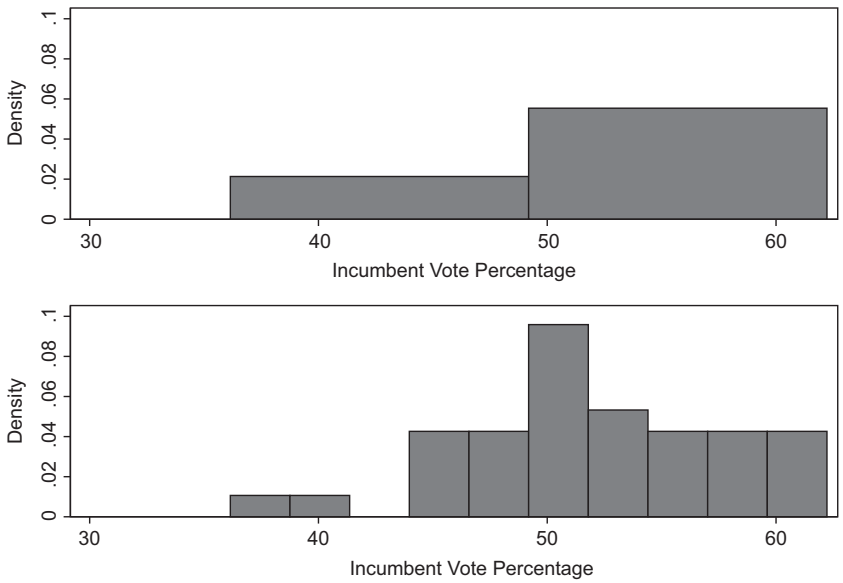

Figure 6.6 Histograms of incumbent-party presidential vote percentage, 1876–2016, depicted with two and then ten blocks

dimension (*x* axis) displays the values of the variable and the vertical dimension (*y* axis) displays the relative frequency of cases. One of the most popular visual depictions of a variable's distribution is the **histogram**, such as Figure 6.5. One problem with histograms is that we (or the computer program with which we are working) must choose how many rectangular blocks (called "bins") are depicted in our histogram. Changing the number of blocks in a histogram can change our impression of the distribution

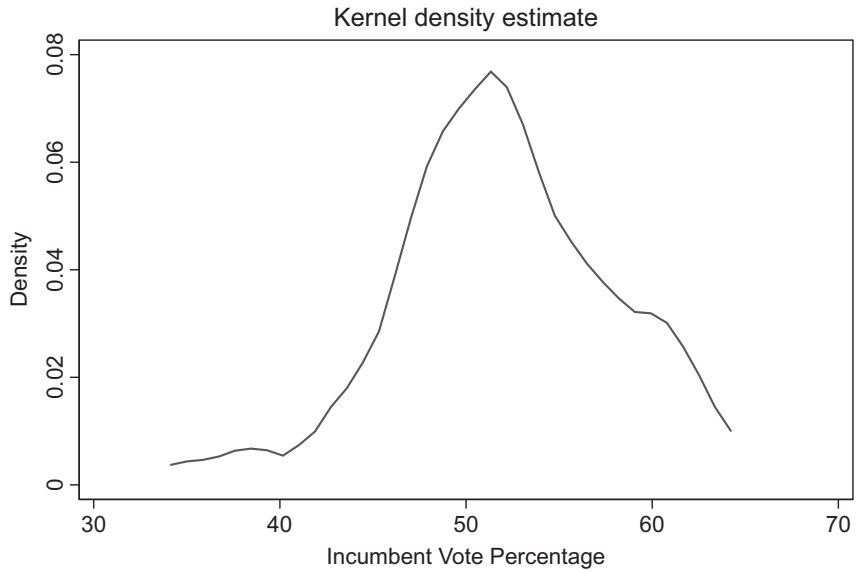

Figure 6.7 Kernel density plot of incumbent-party presidential vote percentage, 1876–2016

of the variable being depicted. Figure 6.6 shows the same variable as in Figure 6.5 with two and then ten blocks. Although we generate both of the graphs in Figure 6.6 from the same data, they are fairly different from each other.

Another option is the **kernel density plot**, as in Figure 6.7, which is based on a smoothed calculation of the density of cases across the range of values.

6.5 LIMITATIONS OF DESCRIPTIVE STATISTICS AND GRAPHS

The tools that we have presented in the last three sections of this chapter are helpful for providing a first look at data, one variable at a time. Taking a look at your data with these tools will help you to better know your data and make fewer mistakes in the long run. It is important, however, to note that we cannot test causal theories with a single variable. After all, as we have noted, a theory is a tentative statement about the possible causal relationship between two variables. Because we have discussed how to describe only a single variable, we have not yet begun to subject our causal theories to appropriate tests.

6.6 CONCLUSIONS

The tools that we have presented in this chapter are helpful for providing a first look at data, one variable at a time. Taking a look at your data with these tools will help you to better know your data and make less mistakes

in the long run. It is important, however, to note that we cannot test causal theories with a single variable. After all, as we have noted, a theory is a tentative statement about the possible causal relationship between two variables. Since we have only discussed how to describe a single variable, we have not yet begun to subject our causal theories to appropriate tests.

CONCEPTS INTRODUCED IN THIS CHAPTER

- categorical variable – a variable for which cases have values that are either different from or the same as the values for other cases, but about which we cannot make any universally holding ranking distinctions
- central tendency – typical values for a particular variable at the center of its distribution
- continuous variable – a variable whose metric has equal unit differences such that a one-unit increase in the value of the variable indicates the same amount of change across all values of that variable
- dispersion – the spread or range of values of a variable
- equal unit differences – a variable has equal unit differences if a one-unit increase in the value of that variable always means the same thing
- expected value – a synonym for mean value
- histogram – a visual depiction of the distribution of a single variable that produces a two-dimensional figure in which the horizontal dimension (x axis) displays the values of the variable and the vertical dimension (y axis) displays the relative frequency of cases
- kernel density plot – a visual depiction of the distribution of a single variable based on a smoothed calculation of the density of cases across the range of values
- kurtosis – a statistical measure indicating the steepness of the statistical distribution of a single variable
- least-squares property – a property of the mean value for a single variable Y, which means that the sum of the squared differences between each Y value, Y_i, and the mean value of Y, $\bar{Y}$, is less than the sum of the squared differences between each Y value, Y_i, and some value c, for all ($\forall$) c values not equal to ($\neq$) $\bar{Y}$
- mean value – the arithmetical average of a variable equal to the sum of all values of Y across individual cases of Y, Y_i, divided by the total number of cases
- measurement metric – the type of values that the variable takes on
- median value – the value of the case that sits at the exact center of our cases when we rank the values of a single variable from the smallest to the largest observed values

- mode – the most frequently occurring value of a variable
- ordinal variable – a variable for which we can make universally holding ranking distinctions across the variable values, but whose metric does not have equal unit differences
- outlier – a case for which the value of the variable is extremely high or low relative to the rest of the values for that variable
- rank statistics – a class of statistics used to describe the variation of continuous variables based on their ranking from lowest to highest observed values
- skewness – a statistical measure indicating the symmetry of the distribution around the mean
- standard deviation – a statistical measure of the dispersion of a variable around its mean
- statistical moments – a class of statistics used to describe the variation of continuous variables based on numerical calculations
- variance – a statistical measure of the dispersion of a variable around its mean
- variation – the distribution of values that a variable takes across the cases for which it is measured
- zero-sum property – a property of the mean value for a single variable Y, which means that the sum of the difference between each Y value, Y_i, and the mean value of Y, $\bar{Y}$, is equal to zero

EXERCISES

1. *Collecting and describing a categorical variable*. Find data for a categorical variable in which you are interested. Get those data into a format that can be read by the statistical software that you are using. Produce a frequency table and describe what you see.
2. *Collecting and describing a continuous variable*. Find data for a continuous variable in which you are interested. Get those data into a format that can be read by the statistical software that you are using. Produce a table of descriptive statistics and either a histogram or a kernel density plot. Describe what you have found out from doing this.
3. In Table 6.1, why would it be problematic to calculate the mean value of the variable "Religious Identification"?
4. *Moving from mathematical formulae to textual statements*. Write a sentence that conveys what is going on in each of the following equations:

   (**a**) $Y = 3 \; \forall \; X_i = 2$.

   (**b**) $Y_{\text{total}} = \sum_{i=1}^{n} Y_i = n\bar{Y}$.