5 Getting to Know Your Data: Evaluating Measurement and Variations

OVERVIEW

Although what political scientists care about is discovering whether causal relationships exist between concepts, what we *actually* examine is statistical associations between variables. Therefore it is critical that we have a clear understanding of the concepts that we care about so we can measure them in a valid and reliable way. In this chapter we focus on two critical tasks in the process of evaluating causal theories: measurement and descriptive statistics. As we discuss the importance of measurement, we use several examples from the political science literature, such as the concept of political tolerance. We know that political tolerance and intolerance is a "real" thing – that it exists to varying degrees in the hearts and minds of people. But how do we go about measuring it? What are the implications of poor measurement? Descriptive statistics and descriptive graphs, which represent the second focus of this chapter, 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.

I know it when I see it.

– Associate Justice of the United States Supreme Court Potter Stewart, in an attempt to define "obscenity" in a concurring opinion in *Jacobellis v. Ohio* (1964)

These go to eleven.

– Nigel Tufnel (played by Christopher Guest), describing the volume knob on his amplifier, in the movie *This Is Spinal Tap*

5.1 GETTING TO KNOW YOUR DATA

We have emphasized the role of theory in political science. That is, we care about causal relationships between concepts that interest us as political scientists. At this point, you are hopefully starting to develop theories of your own about politics. If these original theories are in line with the rules of the road that we laid out in Chapter 1, they will be causal, general, and parsimonious. They may even be elegant and clever.

But at this point, it is worth pausing and thinking about what a theory really *is* and *is not*. To help us in this process, take a look back at Figure 1.2. A theory, as we have said, is merely a conjecture about the possible causal relationship between two or more concepts. As scientists, we must always resist the temptation to view our theories as somehow supported until we have evaluated evidence from the real world, and until we have done everything we can with empirical evidence to evaluate how well our theory clears the four causal hurdles we identified in Chapter 3. In other words, we cannot evaluate a theory until we have gone through the rest of the process depicted in Figure 1.2. The first part of this chapter deals with operationalization, or the movement of variables from the rather abstract conceptual level to the very real measured level. We can conduct hypothesis tests and make reasonable evaluations of our theories only after we have gone carefully through this important process with all of our variables.

If our theories are statements about relationships *between concepts*, when we look for evidence to test our theories, we are immediately confronted with the reality that we do not actually *observe* those concepts. Many of the concepts that we care about in political science, as we will see shortly, are inherently elusive and downright impossible to observe empirically in a direct way, and sometimes incredibly difficult to measure quantitatively. For this reason, we need to think very carefully about the data that we choose to evaluate our theories.

Until now, we have seen many examples of data, but we have not discussed the process of obtaining data and putting them to work. If we think back to Figure 1.2, we are now at the stage where we want to move from the theoretical-conceptual level to the empirical-measured level. For every theoretical concept, there are multiple operationalization or measurement strategies. As we discussed in the previous chapter, one of the first major decisions that one needs to make is whether to conduct an experiment or some form of observational test. In this chapter, we assume that you have a theory and that you are going to conduct an observational test of your theory.

A useful exercise, once you have developed an original theory, is to draw your version of Figure 1.2 and to think about what would be the ideal setup for testing your theory. What would be the best setup, a cross-sectional design or a time-series design? Once you have answered this question and have your ideal time and spatial dimensions in hand, what would be the ideal measure of your independent and dependent variables?

Having gone through the exercise of thinking about the ideal data, the first instinct of most students is to collect their own data, perhaps even to do so through a survey.[1] In our experience, beginning researchers almost always underestimate the difficulties and the costs (in terms of both time and money) of collecting one's own data. We *strongly* recommend that you look to see what data are already available for you to use.

For a political science researcher, one of the great things about the era in which we live is that there is a nearly endless supply of data that are available from web sites and other easily accessible resources.[2] But a few words of caution: just because data are easily available on the web does not mean that these data will be perfectly suitable to the particular needs of your hypothesis test. What follows in the rest of this chapter is a set of considerations that you should have in mind to help you determine whether or not a particular set of data that you have found is appropriate for your purposes and to help you to get to know your data once you have loaded them into a statistical program. We begin with the all-important topic of variable measurement. We describe the problems of measurement and the importance of measuring the concepts in which we are interested as precisely as possible. During this process, you will learn some thinking skills for evaluating the measurement strategies of scholarship that you read, as well as learn about evaluating the usefulness of measures that you are considering using to test your hypotheses.

We begin the section on measurement in the social sciences generally. We focus on examples from economics and psychology, two social sciences that are at rather different levels of agreement about the measurement of their major variables. In political science, we have a complete range of variables in terms of the levels of agreement about how they should be measured. We discuss the core concepts of measurement and give some examples from political science research. Throughout our discussion of these core concepts, we focus on the measurements of variables that take on a numeric range of

[1] A survey is a particularly cumbersome choice because, at least at most universities, you would need to have approval for conducting your survey from the Human Subjects Research Committee.

[2] One resource that is often overlooked is your school's library. While libraries may seem old-fashioned, your school's library may have purchased access to data sources and librarians are often experts in the location of data from the web.

values we feel comfortable treating the way that we normally treat numeric values. Toward the end of the chapter, when we discuss the basics of getting to know your data with a software program, we will discuss this further and focus on some variable types that can take different types of nonnumeric values.

5.2 SOCIAL SCIENCE MEASUREMENT: THE VARYING CHALLENGES OF QUANTIFYING HUMANITY

Measurement is a "problem" in all sciences – from the physical sciences of physics and chemistry to the social sciences of economics, political science, psychology, and the rest. But in the physical sciences, the problem of measurement is often reduced to a problem of instrumentation, in which scientists develop well-specified protocols for measuring, say, the amount of gas released in a chemical reaction or the amount of light given off by a star. The social sciences, by contrast, are younger sciences, and scientific consensus on how to measure our important concepts is rare. Perhaps more crucial, though, is the fact that the social sciences deal with an inherently difficult-to-predict subject matter: human beings.

The problem of measurement exists in all of the social sciences. It would be wrong, though, to say that it is equally problematic in all of the social science disciplines. Some disciplines pay comparatively little heed to issues of measurement, whereas others are mired nearly constantly in measurement controversies and difficulties.

Consider the subject matter in much research in economics: dollars (or euros, or yen, or what have you). If the concept of interest is "economic output" (or "Gross Domestic Product"), which is commonly defined as the total sum of goods and services produced by labor and property in a given time period, then it is a relatively straightforward matter to obtain an empirical observation that is consistent with the concept of interest.[3] Such measures will not be controversial among the vast majority of scholars. To the contrary, once economists agree on a measure of economic output, they can move on to the next (and more interesting) step in the scientific process – to argue about what forces *cause* greater or less growth in economic output. (That's where the agreement among economists ends.)

Not every concept in economics is measured with such ease, however. Many economists are concerned with poverty: Why are some individuals poor whereas others are not? What forces cause poverty to rise or fall over time? Despite the fact that we all know that poverty is a very real thing,

[3] For details about how the federal government measures GDP, see `http://www.bea.gov`.

measuring who is poor and who is not poor turns out to be a bit tricky. The federal government defines the concept of poverty as "a set of income cutoffs adjusted for household size, the age of the head of the household, and the number of children under age 18."[4] The intent of the cutoffs is to describe "minimally decent levels of consumption."[5] There are difficulties in obtaining empirical observations of poverty, though. Among them, consider the reality that most Western democracies (including the United States) have welfare states that provide transfer payments – in the form of cash payments, food stamps, or services like subsidized health care – to their citizens below some income threshold. Such programs, of course, are designed to minimize or eliminate the problems that afflict the poor. When economists seek to measure a person's income level to determine whether or not he is poor, should they use a "pretransfer" definition of income – a person's or family's income level *before* receiving any transfer payments from the government – or a "posttransfer" definition? Either choice carries some negative consequences. Choosing a pretransfer definition of income gives a sense of how much the private sector of the economy is failing. On the other hand, a posttransfer definition gives a sense of how much welfare state programs are falling short and how people are actually living. As the Baby Boom generation in the United States continues to age more and more people are retiring from work. Using a pretransfer measure of poverty means that researchers will not consider Social Security payments – the U.S.'s largest source of transfer payments by far – and therefore the (pretransfer) poverty rate should grow rather steadily over the next few decades, regardless of the health of the overall economy. This might not accurately represent what we mean by "poverty" (Danziger and Gottschalk 1983).

If, owing to their subject matter, economists rarely (but occasionally) have measurement obstacles, the opposite end of the spectrum would be the discipline of psychology. The subject matter of psychology – human behavior, cognition, and emotion – is rife with concepts that are extremely difficult to measure. Consider a few examples. We all know that the concept of "depression" is a real thing; some individuals are depressed, and others are not. Some individuals who are depressed today will not be depressed as time passes, and some who are not depressed today will become depressed. Yet how is it possible to assess scientifically whether a person is or is not

[4] See `http://www.census.gov/hhes/www/poverty/poverty.html`.

[5] Note a problem right off the bat: What is "minimally decent"? Do you suspect that what qualified as "minimally decent" in 1950 or 1985 would be considered "minimally decent" today? This immediately raises issues of how sensible it is to compare the poverty rates from the past with those of today. If the floor of what is considered minimally decent continues to rise, then the comparison is problematic at best, and meaningless at worst.

depressed?[6] Why does it matter if we measure depression accurately? Recall the scientific stakes described at the beginning of this chapter: If we don't measure depression well, how can we know whether remedies like clinical therapy or chemical antidepressants are effective?[7] Psychology deals with a variety of other concepts that are notoriously slippery, such as the clinical focus on "anxiety," or the social-psychological focus on concepts such as "stereotyping" or "prejudice" (which are also of concern to political scientists).

Political science, in our view, lies somewhere between the extremes of economics and psychology in terms of how frequently we encounter serious measurement problems. Some subfields in political science operate relatively free of measurement problems. The study of political economy – which examines the relationship between the economy and political forces such as government policy, elections, and consumer confidence – has much the same feel as economics, for obvious reasons. Other subfields encounter measurement problems regularly. The subfield of political psychology – which studies the way that individual citizens interact with the political world – shares much of the same subject matter as social psychology, and hence, because of its focus on the attitudes and feelings of people, it shares much of social psychology's measurement troubles.

Consider the following list of critically important concepts in the discipline of political science that have sticky measurement issues:

- **Judicial activism**: In the United States, the role of the judiciary in the policy-making process has always been controversial. Some view the federal courts as the protectors of important civil liberties, whereas others view the courts as a threat to democracy, because judges are not elected. How is it possible to identify an "activist judge" or an "activist decision"?[8]
- **Congressional roll-call liberalism**: With each successive session of the U.S. Congress, commentators often compare the level of liberalism and

[6] Since 1952, the American Psychiatric Press, Inc., has published the *Diagnostic and Statistical Manual of Mental Disorders*, now in its fifth edition (called DSM 5), which diagnoses depression by focusing on four sets of symptoms that indicate depression: mood, behavioral symptoms such as withdrawal, cognitive symptoms such as the inability to concentrate, and somatic symptoms such as insomnia.

[7] In fact, the effectiveness of clinical "talk" therapy is a matter of some contention among psychologists. See "Married with Problems? Therapy May Not Help," *New York Times*, April 19, 2005.

[8] In this particular case, there could even be a disagreement over the conceptual definition of "activist." What a conservative and a liberal would consider to be "activist" might produce no agreement at all. See "Activist, Schmactivist," *New York Times*, August 15, 2004, for a journalistic account of this issue.

conservatism of the present Congress with that of its most recent predecessors. How do we know if the Congress is becoming more or less liberal over time (Poole and Rosenthal 1997)?

- **Political legitimacy**: How can analysts distinguish between a "legitimate" and an "illegitimate" government? The key conceptual issue is more or less "how citizens evaluate governmental authority" (Weatherford 1992). Some view it positively, others quite negatively. Is legitimacy something that can objectively be determined, or is it an inherently subjective property among citizens?
- **Political sophistication**: Some citizens know more about politics and are better able to process political information than other citizens who seem to know little and care less about political affairs. How do we distinguish politically sophisticated citizens from the politically unsophisticated ones? Moreover, how can we tell if a society's level of political sophistication is rising or falling over time (Luskin 1987)?
- **Social capital**: Some societies are characterized by relatively high levels of interconnectedness, with dense networks of relationships that make the population cohesive. Other societies, in contrast, are characterized by high degrees of isolation and distrustfulness. How can we measure what social scientists call *social capital* in a way that enables us to compare one society's level of connectedness with another's or one society's level of connectedness at varying points in time (Putnam 2000)?

In Sections 5.4 and 5.5, we describe the measurement controversies surrounding two other concepts that are important to political science – democracy and political tolerance. But first, in the next section, we describe some key issues that political scientists need to grapple with when measuring their concepts of interest.

5.3 PROBLEMS IN MEASURING CONCEPTS OF INTEREST

We can summarize the problems of measuring concepts of interest in preparation for hypothesis testing as follows: First, you need to make sure that you have conceptual clarity. Next, settle on a reasonable level of measurement. Finally, ensure that your measure is both valid and reliable. After you repeat this process for each variable in your theory, you are ready to test your hypothesis.

Unfortunately, there is no clear map to follow as we go through these steps with our variables. Some variables are very easy to measure, whereas others, because of the nature of what we are trying to measure, will always be elusive. As we will see, debates over issues of measurement are at the core of many interesting fields of study in political science.

5.3.1 Conceptual Clarity

The first step in measuring any phenomenon of interest to political scientists is to have a clear sense of what the concept is that we are trying to measure. In some cases, like the ones we subsequently discuss, this is an exceedingly revealing and difficult task. It requires considerably disciplined thought to ferret out precisely what we mean by the concepts about which we are theorizing. But even in some seemingly easy examples, this is more difficult than might appear at first glance.

Consider a survey in which we needed to measure a person's *income*. That would seem easy enough. Once we draw our sample of adults, why not just ask each respondent, "What is your income?" and offer a range of values, perhaps in increments of $10,000 or so, on which respondents could place themselves. What could be the problem with such a measure? Imagine a 19-year-old college student whose parents are very wealthy, but who has never worked herself, answering such a question. How much income has that person earned in the last year? Zero. In such a circumstance, this is the true answer to such a question. But it is not a particularly valid measure of her income. We likely want a measure of income that reflects the fact that her parents earn a good deal of money, which affords her the luxury of not having to work her way through school as many other students do. That measure should place the daughter of wealthy parents ahead of a relatively poor student who carries a full load and works 40 hours a week just to pay her tuition. Therefore, we might reconsider our seemingly simple question and ask instead, "What is the total amount of income earned in the most recently completed tax year by you and any other adults in your household, including all sources of income?" This measure puts the nonworking child of wealthy parents ahead of the student from the less-well-off family. And, for most social science purposes, this is the measure of "income" that we would find most theoretically useful.[9]

At this point, it is worth highlighting that the *best* measure of income – as well as that of most other concepts – depends on what our theoretical objectives are. The best measure of something as simple as a respondent's income depends on what we intend to relate that measure to in our hypothesis testing.

5.3.2 Reliability

An operational measure of a concept is said to be reliable to the extent that it is repeatable or consistent; that is, applying the same measurement

[9] The same issues would arise in assessing the income of retired people who no longer participate in the workforce.

rules to the same case or observation will produce identical results. An unreliable measure, by contrast, would produce inconsistent results for the same observation. For obvious reasons, all scientists want their measures to be reliable.

Perhaps the most simple example to help you understand this is your bathroom scale. Say you step up on the scale one morning and the scale tells you that you weigh 150 pounds. You step down off the scale and it returns to zero. But have you ever *not* trusted that scale reading, and thought to yourself, "Maybe if I hop back up on the scale, I'll get a number I like better?" That is a **reliability** check. If you (immediately) step back on the scale, and it tells you that you now weigh 146 pounds, your scale is unreliable, because repeated measures of the same case – your body at that particular point in time – produced different results.

To take our bathroom scale example to the extreme, we should not confuse over-time variability with unreliability. If you wake up 1 week later and weigh 157 instead of 150 that does not necessarily mean that your scale is unreliable (though that might be true). Perhaps you substituted french fries for salads at dinner in the intervening week, and perhaps you exercised less vigorously or less often.

Reliability is often an important issue when scholars need to code events or text for quantitative analysis. For example, if a researcher was trying to code the text of news coverage that was favorable or unfavorable toward a candidate for office, he would develop some specific coding rules to apply to the text – in effect, to count certain references as either "pro" or "con" with respect to the candidate. Suppose that, for the coding, the researcher employs a group of students to code the text – a practice that is common in political research. A *reliable* set of coding rules would imply that, when one student applies the rules to the text, the results would be the same as when another student takes the rules and applies them to the same text. An *unreliable* set of coding rules would imply the opposite, namely, that when two different coders try to apply the same rules to the same news articles, they reach different conclusions.[10] The same issues arise when one codes things such as events by using newspaper coverage.[11]

5.3.3 Measurement Bias and Reliability

One of the concerns that comes up with any measurement technique is **measurement bias**, which is the systematic over-reporting or under-reporting of

[10] Of course, it is possible that the coding *scheme* is perfectly reliable, but the *coders themselves* are not.

[11] There are a variety of tools for assessing reliability, many of which are beyond the scope of this discussion.

values for a variable. Although measurement bias is a serious problem for anyone who wants to know the "true" values of variables for particular cases, it is less of a problem than you might think for theory-testing purposes. To better understand this, imagine that we have to choose between two different operationalizations of the same variable. Operationalization A is biased but reliable, and Operationalization B is unbiased but unreliable. For theory-testing purposes we would greatly prefer the biased but reliable Operationalization A!

You will be better able to see why this is the case once you have an understanding of statistical hypothesis testing from Chapters 7 and beyond. For now, though, keep in mind that as we test our theories we are looking for general patterns between two variables. For instance, with *higher* values of X do we tend to see *higher* values of Y, or with *higher* values of X do we tend to see *lower* values of Y? If the measurement of X was biased upward, the same general pattern of association with Y would be visible. But if the measurement of X was unreliable, it would obscure the underlying relationship between X and Y.

5.3.4 Validity

The most important feature of a measure is that it is valid. A valid measure accurately represents the concept that it is supposed to measure, whereas an invalid measure measures something other than what was originally intended. All of this might sound a bit circular, we realize.

Perhaps it is useful to think of some important concepts that represent thorny measurement examples in the social sciences. In both social psychology and political science, the study of the concept of *prejudice* has been particularly important. Among individuals, the level of prejudice can vary, from vanishingly small amounts to very high levels. Measuring prejudice can be important in social–psychological terms, so we can try to determine what factors cause some people to be prejudiced whereas others do not. In political science, in particular, we are often interested in the attitudinal and behavioral consequences of prejudice. Assuming that some form of truth serum is unavailable, how can we obtain a quantitative measure of prejudice that can tell us who harbors large amounts of prejudice, who harbors some, and who harbors none? It would be easy enough to ask respondents to a survey if they were prejudiced or not. For example, we could ask respondents: "With respect to people who have a different race or ethnicity than you, would you say that you are extremely prejudiced, somewhat prejudiced, mildly prejudiced, or not at all prejudiced toward them?" But we would have clear reasons to doubt the **validity** of their answers – whether

their measured responses accurately reflected their true levels of prejudice.

There are a variety of ways to assess a measure's validity, though it is critical to note that all of them are theoretical and subject to large degrees of disagreement. There is no simple formula to check for a measure's validity on a scale of 0 to 100, unfortunately. Instead, we rely on several overlapping ways to determine a measure's validity. First, and most simply, we can examine a measure's **face validity**. When examining a measurement strategy, we can first ask whether or not, on its face, the measure appears to be measuring what it purports to be measuring. This is face validity. Second, and a bit more advanced, we can scrutinize a measure's **content validity**. What is the concept to be measured? What are all of the essential elements to that concept and the features that define it? And have you excluded all of the things that are not it? For example, the concept of democracy surely contains the element of "elections," but it also must incorporate more than mere elections, because elections are held in places like North Korea, which we know to be nondemocratic. What else must be in a valid measure of democracy? (More on this notion later on.) Basically, content validation is a rigorous process that forces the researcher to come up with a list of all of the critical elements that, as a group, define the concept we wish to measure. Finally, we can examine a measure's **construct validity**: the degree to which the measure is related to other measures that theory requires them to be related to. That is, if we have a theory that connects democratization and economic development, then a measure of democracy that is related to a measure of economic development (as our theory requires) serves simultaneously to confirm the theory and also to validate the measure of democracy. Of course, one difficulty with this approach is what happens when the expected association is not present. Is it because our measure of democracy is invalid or because the theory is misguided? There is no conclusive way to tell.

5.3.5 The Relationship between Validity and Reliability

What is the connection between validity and reliability? Is it possible to have a valid but unreliable measure? And is it possible to have a reliable but invalid measure? With respect to the second question, some scientific debate exists; there are some who believe that it is possible to have a reliable but invalid measure. In our view, that is possible in abstract terms. But because we are interested in measuring concepts in the interest of evaluating causal theories, we believe that, in all practical terms, any conceivable measures that are reliable but invalid will not be useful in evaluating causal theories.

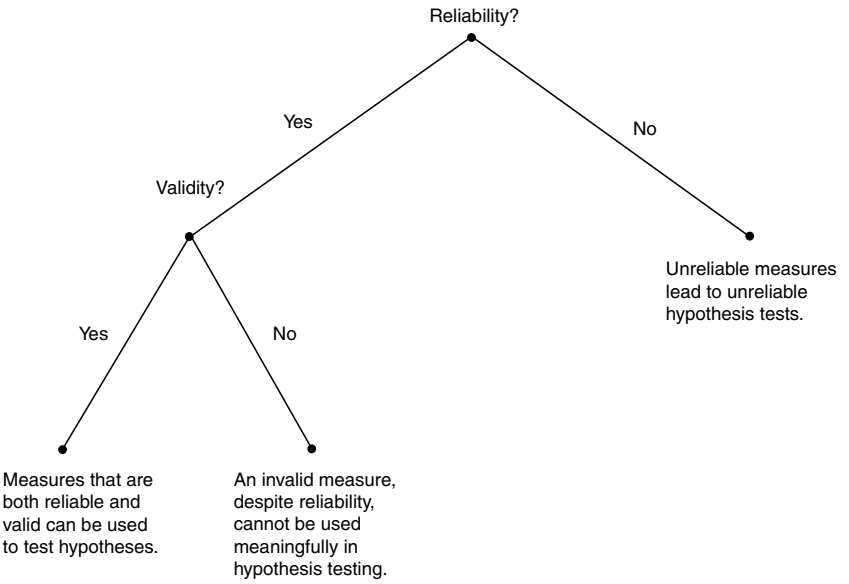

Figure 5.1. Reliability, validity, and hypothesis testing.

Similarly, it is theoretically possible to have valid but unreliable measures. But those measures also will be problematic for evaluating causal theories, because we will have no confidence in the hypothesis tests that we conduct. We present the relationship between reliability and validity in Figure 5.1, where we show that, if a measure is unreliable, there is little point in evaluating its validity. Once we have established that a measure is reliable, we can assess its validity, and only reliable and valid measures are useful for evaluating causal theories.

5.4 CONTROVERSY 1: MEASURING DEMOCRACY

Although we might be tempted to think of democracy as being similar to pregnancy – that is, a country either *is* or *is not* a democracy much the same way that a woman either *is* or *is not* pregnant – on a bit of additional thought, we are probably better off thinking of democracy as a *continuum*.[12] That is, there can be varying degrees to which a government is democratic. Furthermore, within democracies, some countries are more democratic than others, and a country can become more or less democratic as time passes.

[12] This position, though, is controversial within political science. For an interesting discussion about whether researchers should measure democracy as a binary concept or a continuous one, see Elkins (2000).

But defining a continuum that ranges from democracy, on one end, to totalitarianism, on the other end, is not at all easy. We might be tempted to resort to the Potter Stewart "I know it when I see it" definition. As political scientists, of course, this is not an option. We have to begin by asking ourselves, what do we mean by democracy? What are the core elements that make a government more or less democratic? Political philosopher Robert Dahl (1971) persuasively argued that there are two core attributes to a democracy: "contestation" and "participation." That is, according to Dahl, democracies have competitive elections to choose leaders and broadly inclusive rules for and rates of participation.

Several groups of political scientists have attempted to measure democracy systematically in recent decades.[13] The best known – though by no means universally accepted – of these is the Polity IV measure.[14] The project measures democracy with annual scores ranging from −10 (strongly autocratic) to +10 (strongly democratic) for every country on Earth from 1800 to 2004.[15] In these researchers' operationalization, democracy has four components:

1. Regulation of executive recruitment
2. Competitiveness of executive recruitment
3. Openness of executive recruitment
4. Constraints on chief executive

For each of these dimensions, experts rate each country on a particular scale. For example, the first criterion, "regulation of executive recruitment," allows for the following possible values:

- +3 = regular competition between recognized groups
- +2 = transitional competition
- +1 = factional or restricted patterns of competition
- 0 = no competition

Countries that have regular elections between groups that are more than ethnic rivals will have higher scores. By similar procedures, the scholars associated with the project score the other dimensions that comprise their democracy scale.

[13] For a useful review and comparison of these various measures, see Munck and Verkuilen (2002).

[14] The project's web site, which provides access to a vast array of country-specific over-time data, is `http://www.cidcm.umd.edu/inscr/polity`.

[15] They derive the scores on this scale from two separate 10-point scales, one for democracy and the other for autocracy. A country's Polity score for that year is its democracy score minus its autocracy score; thus, a country that received a 10 on the democracy scale and a 0 on the autocracy scale would have a net Polity score of 10 for that year.

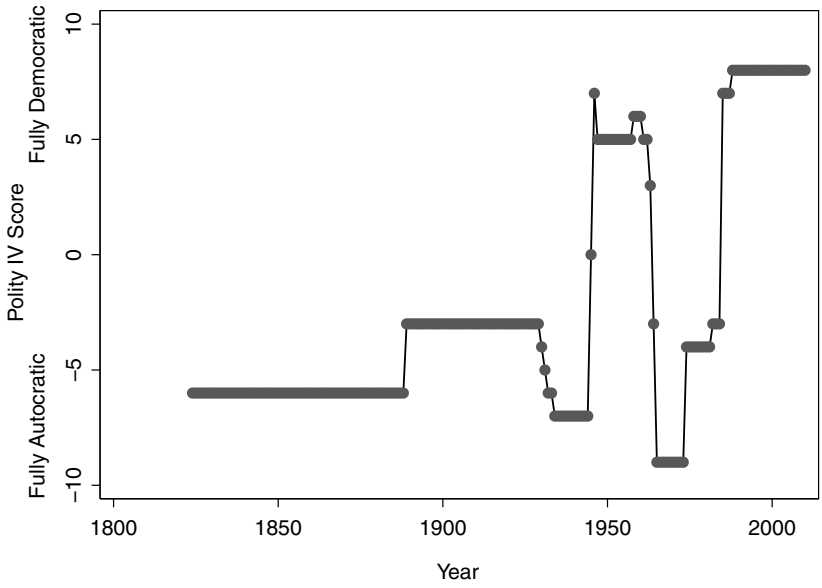

Figure 5.2. Polity IV score for Brazil.

Figure 5.2 presents the Polity score for Brazil from 1824 through 2010.[16] Remember that higher scores represent points in time when Brazil was more democratic, and lower scores represent times when Brazil was more autocratic. There has been, as you can see, enormous **variation** in the democratic experience in Brazil since its declaration of independence from Portugal in 1822. If we make a rough comparison of these scores with the timeline of Brazil's political history, we can get an initial evaluation of the face validity of the Polity scores as a measure of democracy. After the declaration of independence from Portugal, Brazil was a constitutional monarchy headed by an emperor. After a coup in 1889, Brazil became a republic, but one in which politics was fairly strictly controlled by the elites from the two dominant states. We can see that this regime shift resulted in a move from a Polity score of –6 to a score of –3. Starting in 1930, Brazil went through a series of coups and counter-coups. Scholars writing about this period (e.g., Skidmore 2009) generally agree that the nation's government became more and more autocratic during this era. The Polity scores certainly reflect this movement. In 1945, after another military coup, a relatively democratic government was put into place. This regime lasted until the mid 1960s when another period of instability was ended by a military dictatorship. This period is widely recognized as the most politically repressive regime in Brazil's independent political history. It lasted until

[16] Source: `http://www.systemicpeace.org/inscr/inscr.htm`.

1974 when the ruling military government began to allow limited political elections and other political activities. In 1985, Brazil elected a civilian president, a move widely seen as the start of the current democratic period. Each of these major moves in Brazil's political history is reflected in the Polity scores. So, from this rough evaluation, Polity scores have face validity.

The Polity measure is rich in historical detail, as is obvious from Figure 5.2. The coding rules are transparent and clear, and the amount of raw information that goes into a country's score for any given year is impressive. And yet it is fair to criticize the Polity measure for including only one part of Dahl's definition of democracy. The Polity measure contains rich information about what Dahl calls "contestation" – whether a country has broadly open contests to decide on its leadership. But the measure is much less rich when it comes to gauging a country's level of what Dahl calls "participation" – the degree to which citizens are engaged in political processes and activities. This may be understandable, in part, because of the impressive time scope of the study. After all, in 1800 (when the Polity time series begins), very few countries had broad electoral participation. Since the end of World War II, broadly democratic participation has spread rapidly across the globe. But if the world is becoming a more democratic place, owing to expansion of suffrage, our measures of democracy ought to incorporate that reality. Because the Polity measure includes one part ("contestation") of what it means, conceptually, to be democratic, but ignores the other part ("participation"), the measure can be said to lack content validity. The Polity IV measure, despite its considerable strengths, does not fully encompass what it means, conceptually, to be more or less democratic.

This problem is nicely illustrated by examining the Polity score for the United States presented in Figure 5.3, which shows its score for the time period 1800–2010. The consistent score of 10 for almost every year after the founding of the republic – the exception is during the Civil War, when President Lincoln suspended the writ of habeas corpus – belies the fact that the United States, in many important ways, has become a more democratic nation over its history, particularly on the participatory dimension not captured in the Polity measure. Even considering something as basic to democratic participation as the right to vote reveals this to be the case. Slavery prevented African Americans from many things, voting included, until after the Civil War, and Jim Crow laws in the South kept those prohibitions in place for nearly a century afterward. Women, too, were not allowed to vote until the 19th Amendment to the Constitution was ratified in 1920. It would be difficult to argue that these changes did not make the United States more democratic, but of course those changes are not reflected in Figure 5.3. This is not to say that the Polity measure is useless,

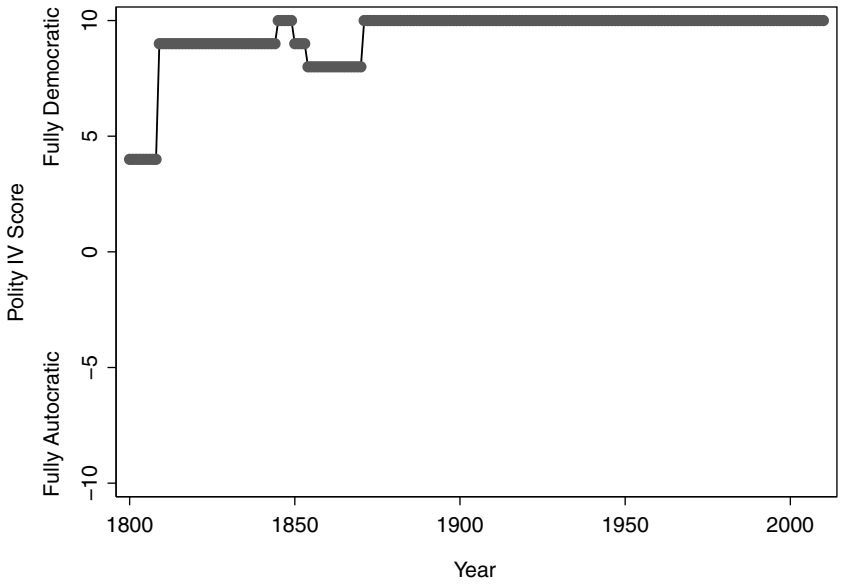

Figure 5.3. Polity IV score for the United States.

but merely that it lacks content validity because one of the key components of democracy – participation – is nowhere to be found in the measure.

5.5 CONTROVERSY 2: MEASURING POLITICAL TOLERANCE

We know that some continuum exists in which, on the one end, some individuals are extremely "tolerant" and, on the other end, other individuals are extremely "intolerant." In other words, political tolerance and intolerance, at the conceptual level, are real things. Some individuals have more tolerance and others have less. It is easy to imagine why political scientists would be interested in political tolerance and intolerance. Are there systematic factors that cause some people to be tolerant and others to be intolerant?

Measuring political tolerance, on the other hand, is far from easy. Tolerance is not like cholesterol, for which a simple blood test can tell us how much of the good and how much of the bad we have inside of us. The naive approach to measuring political tolerance – conducting a survey and asking people directly "Are you tolerant or intolerant?" – seems silly right off the bat. Any such survey question would surely produce extremely high rates of "tolerance," because presumably very few people – even intolerant people – think of themselves as intolerant. Even those who are aware of their own intolerance are unlikely to admit that fact to a pollster. Given this situation, how have political scientists tackled this problem?

During the 1950s, when the spread of Soviet communism represented the biggest threat to America, Samuel Stouffer (1955) conducted a series of opinion surveys to measure how people reacted to the Red Scare. He asked national samples of Americans whether they would be willing to extend certain civil liberties – like being allowed to teach in a public school, to be free from having phones tapped, and the like – to certain unpopular groups like communists, socialists, and atheists. He found that a variety of people were, by these measures, intolerant; they were not willing to grant these civil liberties to members of those groups. The precise amount of intolerance varied, depending on the target group and the activity mentioned in the scenarios, but intolerance was substantial – at least 70% of respondents gave the intolerant response. Stouffer found that the best predictor of an individual's level of tolerance was how much formal education he or she had received; people with more education emerged as more tolerant, and people with less education were less tolerant. In the 1970s, when the Red Scare was subsiding somewhat, a new group of researchers asked the identical questions to a new sample of Americans. They found that the levels of intolerance had dropped considerably over the 20-odd years – in only one scenario did intolerance exceed 60% and in the majority of scenarios it was below 50% – leading some to speculate that political intolerance was waning.

However, also in the late 1970s, a different group of researchers led by political scientist John Sullivan questioned the *validity* of the Stouffer measures and hence questioned the conclusions that Stouffer reached. The concept of political tolerance, wrote Sullivan, Pierson, and Marcus (1979), "presupposes opposition." That is, unless a survey respondent actively opposed communists, socialists, and atheists, the issue of tolerance or intolerance simply does not arise. By way of example, consider asking such questions of an atheist. Is an atheist who agrees that atheists should be allowed to teach in public schools politically tolerant? Sullivan and his colleagues thought not.

The authors proposed a new set of survey-based questions that were, in their view, more consistent with a conceptual understanding of tolerance. If, as they defined it, tolerance presupposes opposition, then researchers need to *find out* who the survey respondent opposes; *assuming* that the respondent might oppose a particular group is not a good idea. They identified a variety of groups active in politics at the time – including racist groups, both pro- and anti-abortion groups, and even the Symbionese Liberation Army – and asked respondents which one they disliked the most. They followed this up with questions that looked very much like the Stouffer items, only directed at *the respondent's own* disliked groups instead of the ones Stouffer had picked out for them.

Among other findings, two stood out. First, the levels of intolerance were strikingly high. As many as 66% of Americans were willing to forbid members of their least-liked group from holding rallies, and fully 71% were willing to have the government ban the group altogether. Second, under this new conceptualization and measurement of tolerance, the authors found that an individual's perception of the threatening nature of the target group, and not their level of education, was the primary predictor of intolerance. In other words, individuals who found their target group to be particularly threatening were most likely to be intolerant, whereas those who found their most-disliked group to be less threatening were more tolerant. Education did not directly affect tolerance either way. In this sense, measuring an important concept differently produced rather different substantive findings about causes and effects.[17]

It is important that you see the connection to valid measurement here. Sullivan and his colleagues argued that Stouffer's survey questions were not valid measures of tolerance because the question wording did not accurately capture what it meant, in the abstract, to be intolerant (specifically, opposition). Creating measures of tolerance and intolerance that more truthfully mirrored the concept of interest produced significantly different findings about the persistence of intolerance, as well as about the factors that cause individuals to be tolerant or intolerant.

5.6 ARE THERE CONSEQUENCES TO POOR MEASUREMENT?

What happens when we fail to measure the key concepts in our theory in a way that is both valid and reliable? Refer back to Figure 1.2, which highlights the distinction between the abstract concepts of theoretical interest and the variables we observe in the real world. If the variables that we observe in the real world do not do a good job of mirroring the abstract concepts, then that affects our ability to evaluate conclusively a theory's empirical support. That is, how can we know if our theory is supported if we have done a poor job measuring the key concepts that we observe? If our empirical analysis is based on measures that do not capture the essence of the abstract concepts in our theory, then we are unlikely to have any confidence in the findings themselves.

5.7 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.

[17] But see Gibson (1992).

But once a researcher has collected data and become familiar and satisfied with how it was 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. What do "typical" values for a variable look like? How tightly clustered (or widely dispersed) are the 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 U.S. presidential elections from 1880 to 2008:[18] 50.22, 49.846, 50.414, 48.268, 47.76, 53.171, 60.006, 54.483, 54.708, 51.682, 36.119, 58.244, 58.82, 40.841, 62.458, 54.999, 53.774, 52.37, 44.595, 57.764, 49.913, 61.344, 49.596, 61.789, 48.948, 44.697, 59.17, 53.902, 46.545, 54.736, 50.265, 51.2, 46.311. 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.

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.[19]

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

[18] 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 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.

[19] Many researchers will present this information in an appendix unless there is something particularly noteworthy about the characteristics of one or more of their variables.

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 and **continuous variables**.

5.8 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 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.

5.8.1 Categorical Variables

Categorical variables are variables for which cases have values that are either different 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.

5.8.2 Ordinal Variables

Like categorical variables, **ordinal variables** are also variables for which cases have values that are either different 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?"[20] which we can code as taking on the following values:

1. Republican
2. Independent
3. Democrat

[20] Almost all U.S. respondents put themselves into one of the first three categories. For instance, in 2004, 1,128 of the 1,212 respondents (93.1%) to the postelection 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.

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 categorial variable.

5.8.3 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.[21] Imagine, for instance, a variable labeled "Age in Years." A one-unit increase in this variable *always* indicates an individual who is 1 year older; this is true when we are talking about a

[21] 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.

case with a value of 21 just as it is when we are talking about a case with a value of 55.

5.8.4 Variable Types and Statistical Analyses

As we saw in the preceding subsections, 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.

5.9 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 another column(s). Table 5.1 shows such a table for the variable

Table 5.1 Frequency table for religious identification in the 2004 NES

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	99.9

"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 5.1, the mode of the distribution is "Protestant," because there are more Protestants than there are members of any other single category.

A typical way in which non-statisticians present frequency data is in a pie graph such as Figure 5.4. 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 5.5, are another graphical way to illustrate frequencies of categorical variables. It is worth noting, however, that most of the 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 5.1.

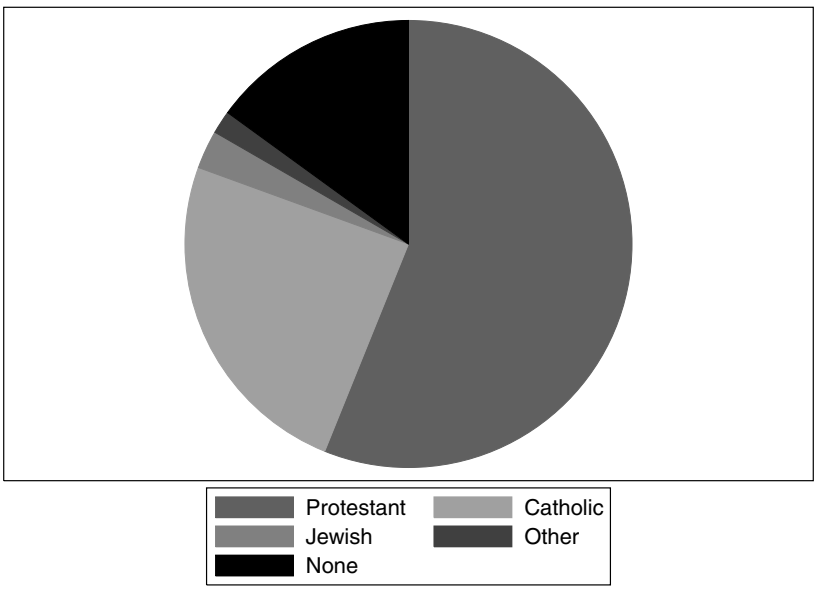

Figure 5.4. Pie graph of religious identification, NES 2004.

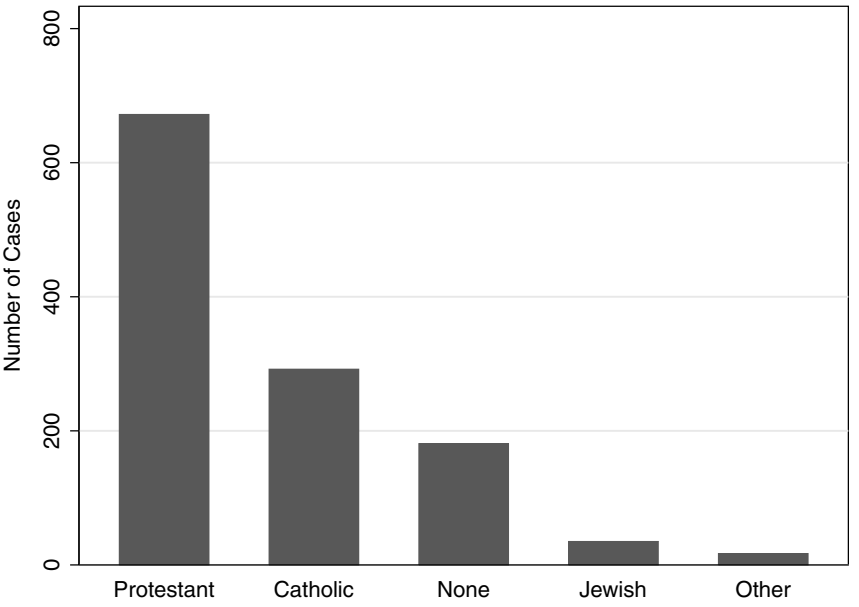

Figure 5.5. Bar graph of religious identification, NES 2004.

5.10 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 5.6 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 U.S. presidential election between 1876 and 2008. The statistics 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

```
. summarize inc_vote, det

                          inc_vote
-------------------------------------------------------------
      Percentiles      Smallest
 1%        36.148         36.148
 5%        40.851         40.851
10%        44.842          44.71       Obs                  34
25%        48.516         44.842       Sum of Wgt.          34

50%       51.4575                      Mean            51.94718
                          Largest      Std. Dev.       5.956539
75%        54.983         60.006
90%        60.006         61.203       Variance        35.48036
95%        61.791         61.791       Skewness       -.3065283
99%        62.226         62.226       Kurtosis        3.100499
```

Figure 5.6. Example output from Stata's "summarize" command with "detail" option.

quite useful together for getting a complete picture of the variation for a single variable.

5.10.1 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 5.6 as the percentage of popular votes for major-party candidates that went to the candidate from the party of the sitting president during U.S. presidential elections from 1876 to 2008. 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 5.2. 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 5.2, we average the value of the two centermost ranked cases to obtain the median value (in our example we calculate the median as $\frac{51.233+51.682}{2} = 51.4575$). This is also known as the value of the variable at the 50% rank. In a similar way, we can talk about the value of the variable at any other percentage rank in which we have an interest. Other ranks that are often of interest

Table 5.2 Values of incumbent vote ranked from smallest to largest

Rank	Year	Value
1	1920	36.148
2	1932	40.851
3	1952	44.71
4	1980	44.842
5	2008	46.311
6	1992	46.379
7	1896	47.76
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	50.22
15	2000	50.262
16	1888	50.414
17	2004	51.233
18	1916	51.682
19	1948	52.319
20	1900	53.171
21	1944	53.778
22	1988	53.832
23	1908	54.483
24	1912	54.708
25	1996	54.737
26	1940	54.983
27	1956	57.094
28	1924	58.263
29	1928	58.756
30	1984	59.123
31	1904	60.006
32	1964	61.203
33	1972	61.791
34	1936	62.226

are the 25% and 75% ranks, which are also known as the first and third "quartile ranks" for a distribution. The difference between the variable value at the 25% and the 75% ranks is known as the "interquartile range" or "IQR" of the variable. In our example variable, the 25% value is 48.516 and the 75% value is 54.983. This makes the IQR = 54.983 – 48.516 = 6.467. 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 5.2, 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% value, they aren't "extremely" higher than the rest of the values. Instead, there seems to be a fairly even progression from the 75% value up to the highest value. The story at the other end of the range of values in Table 5.2 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 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% of the votes cast for the two major parties. A further check of

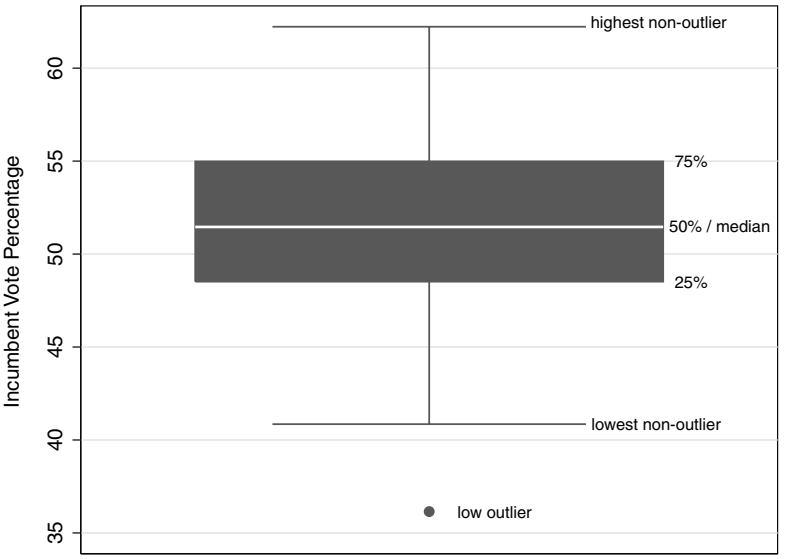

Figure 5.7. Box-whisker plot of incumbent-party presidential vote percentage, 1876–2008.

our data indicates that this was indeed a correct measure of this variable for 1920.[22]

Figure 5.7 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 5.7, we see the median value (or 50% 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% rank and the 75% rank of our variable. The ends of the whiskers show the lowest and highest nonoutlier values of our variable. Each statistical program has its own rules for dealing with outliers, so it is important to know whether 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% value or if its value is 1.5 times the IQR lower than the 25% value. For Figure 5.7 we have set things up

[22] 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.

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 5.2, this is the value of 36.119 from the 1920 election.

5.10.2 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 depicted and calculated as

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

where $\bar{Y}$, known 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.[23] 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**":

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

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's not equal to ($\neq$) $\bar{Y}$. Because of these two properties, the mean value is 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.

[23] 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.

The next statistical moment for a variable is the **variance**. We represent and 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.[24] 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**:

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

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.[25]

The variance and the standard deviation give us a numerical summary of the distribution of cases around the mean value for a variable.[26] We can also visually depict distributions. The idea of visually depicting distributions is to produce 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. One of the most popular visual depictions of a variable's distribution is the **histogram**, such as Figure 5.8.

[24] 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 7.

[25] 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: $(\frac{\sum_{i=1}^{n}|Y_i - \bar{Y}|}{n})$.

[26] The **skewness** and the **excess 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. Both skewness and excess kurtosis are measures that equal zero for the normal distribution, which we will discuss in Chapter 6.

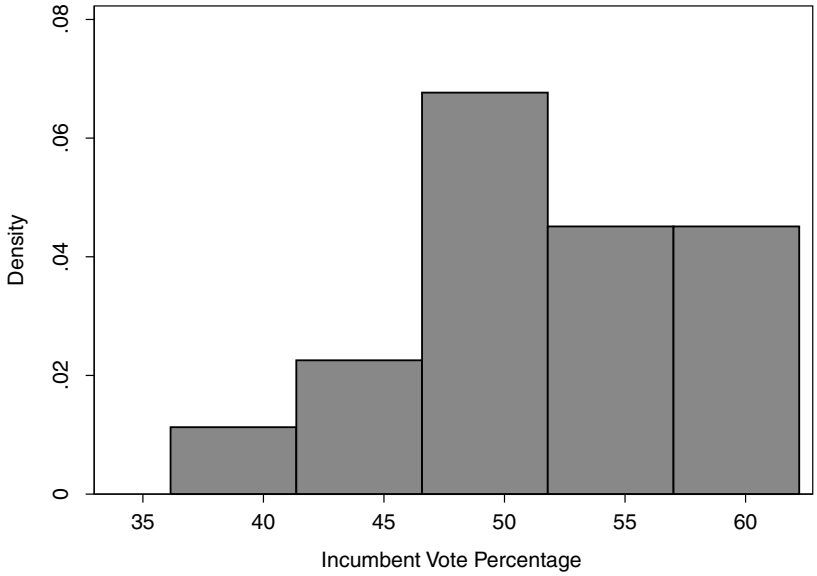

Figure 5.8. Histogram of incumbent-party presidential vote percentage, 1876–2008.

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 of the variable being depicted. Figure 5.9 shows the same variable as in Figure 5.8 with 2 and then 10 blocks. Although we generate both of the graphs in Figure 5.9 from the same data, they are fairly different from each other.

Another option is the **kernel density plot**, as in Figure 5.10, which is based on a smoothed calculation of the density of cases across the range of values.

5.11 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.

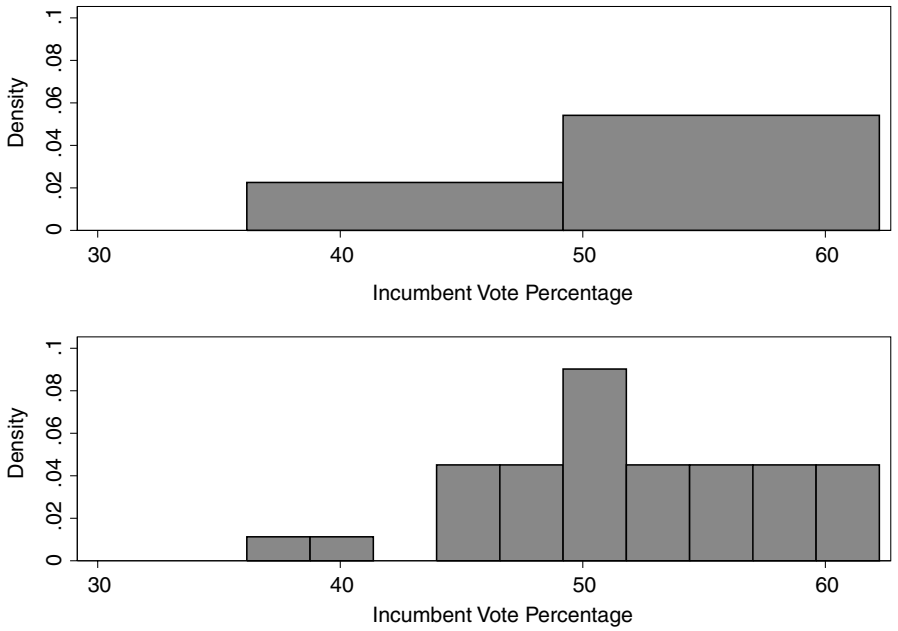

Figure 5.9. Histograms of incumbent-party presidential vote percentage, 1876–2008, depicted with 2 and then 10 blocks.

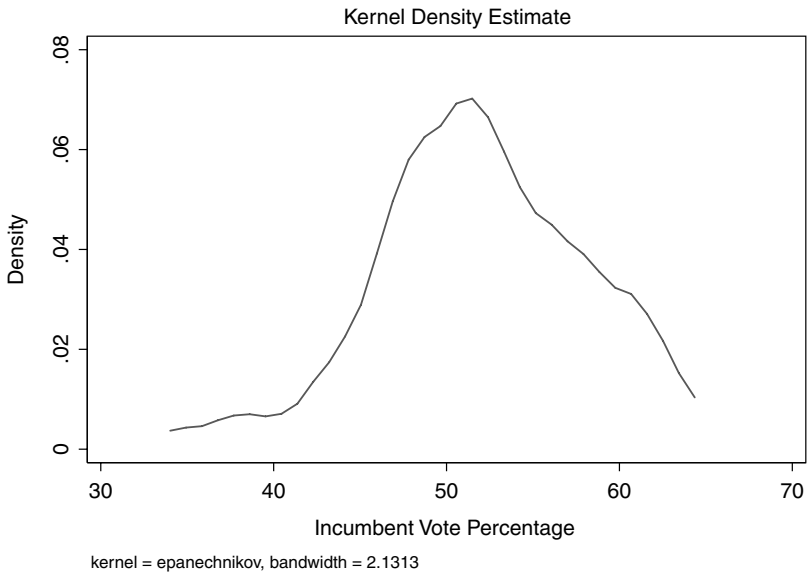

Figure 5.10. Kernel density plot of incumbent-party presidential vote percentage, 1876–2008.

5.12 CONCLUSIONS

How we measure the concepts that we care about matters. As we can see from the preceding examples, different measurement strategies can and sometimes do produce different conclusions about causal relationships.

One of the take-home points of this chapter should be that measurement cannot take place in a theoretical vacuum. The *theoretical purpose* of the scholarly enterprise must inform the process of how we measure what we measure. For example, recall our previous discussion about the various ways to measure poverty. How we want to measure this concept depends on what our objective is. In the process of measuring poverty, if our theoretical aim is to evaluate the effectiveness of different policies at combating poverty, we would have different measurement issues than would scholars whose theoretical aim is to study how being poor influences a person's political attitudes. In the former case, we would give strong consideration to pretransfer measures of poverty, whereas in the latter example, posttransfer measures would likely be more applicable.

The tools that we have presented in this chapter for describing a variable's central tendency and variation 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 variables – variables for which cases have values that are either different 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.
- construct validity – the degree to which the measure is related to other measures that theory requires them to be related to.
- content validity – the degree to which a measure contains all of the critical elements that, as a group, define the concept we wish to measure.

- 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.
- excess kurtosis – a statistical measure indicating the steepness of the statistical distribution of a single variable.
- expected value – a synonym for mean value.
- face validity – whether or not, on its face, the measure appears to be measuring what it purports to be measuring.
- 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.
- 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's 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.
- 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.
- measurement bias – the systematic over-reporting or under-reporting of values for a variable.
- measurement metric – the type of values that the variable takes on.
- 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.

- reliability – the extent to which applying the same measurement rules to the same case or observation will produce identical results.
- 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.
- validity – the degree to which a measure accurately represents the concept that it is supposed to measure.
- 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. Suppose that a researcher wanted to measure the federal government's efforts to make the education of its citizens a priority. The researcher proposed to count the government's budget for education as a percentage of the total GDP and use that as the measure of the government's commitment to education. In terms of validity, what are the strengths and weaknesses of such a measure?
2. Suppose that a researcher wanted to create a measure of media coverage of a candidate for office, and therefore created a set of coding rules to code words in newspaper articles as either "pro" or "con" toward the candidate. Instead of hiring students to implement these rules, however, the researcher used a computer to code the text, by counting the frequency with which certain words were mentioned in a series of articles. What would be the reliability of such a computer-driven measurement strategy, and why?
3. For each of the following concepts, identify whether there would, in measuring the concept, likely be a problem of measurement bias, invalidity, unreliability, or none of the above. Explain your answer.
   (a) Measuring the concept of the public's approval of the president by using a series of survey results asking respondents whether they approve or disapprove of the president's job performance.
   (b) Measuring the concept of political corruption as the percentage of politicians in a country in a year who are convicted of corrupt practices.
   (c) Measuring the concept of democracy in each nation of the world by reading their constitution and seeing if it claims that the nation is "democratic."

Table 5.3. Median incomes of the 50 states, 2004–2005

State	Income	State	Income
Alabama	37,502	Montana	36,202
Alaska	56,398	Nebraska	46,587
Arizona	45,279	Nevada	48,496
Arkansas	36,406	New Hampshire	57,850
California	51,312	New Jersey	60,246
Colorado	51,518	New Mexico	39,916
Connecticut	56,889	New York	46,659
Delaware	50,445	North Carolina	41,820
Florida	42,440	North Dakota	41,362
Georgia	44,140	Ohio	44,349
Hawaii	58,854	Oklahoma	39,292
Idaho	45,009	Oregon	43,262
Illinois	48,008	Pennsylvania	45,941
Indiana	43,091	Rhode Island	49,511
Iowa	45,671	South Carolina	40,107
Kansas	42,233	South Dakota	42,816
Kentucky	36,750	Tennessee	39,376
Louisiana	37,442	Texas	42,102
Maine	43,317	Utah	53,693
Maryland	59,762	Vermont	49,808
Massachusetts	54,888	Virginia	52,383
Michigan	44,801	Washington	51,119
Minnesota	56,098	West Virginia	35,467
Mississippi	34,396	Wisconsin	45,956
Missouri	43,266	Wyoming	45,817

Source: `http://www.census.gov/hhes/www/income/income05/statemhi2.html`. Accessed January 11, 2007.

4. Download a codebook for a political science data set in which you are interested.
   (a) Describe the data set and the purpose for which it was assembled.
   (b) What are the time and space dimensions of the data set?

   Read the details of how one of the variables in which you are interested was coded. Write your answers to the following questions:
   (c) Does this seem like a reliable method of operationalizing this variable? How might the reliability of this operationalization be improved?
   (d) Assess the various elements of the validity for this variable operationalization. How might the validity of this operationalization be improved?

5. If you did not yet do Exercise 5 in Chapter 3, do so now. For the theory that you developed, evaluate the measurement of both the independent and dependent variables. Write about the reliability, and the various aspects of validity for

each measure. Can you think of a better way to operationalize these variables to test your theory?

6. *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.

7. *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.

8. In Table 5.1, why would it be problematic to calculate the mean value of the variable "Religious Identification?"

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

10. *Computing means and standard deviations*. Table 5.3 contains the median income for each of the 50 U.S. states for the years 2004–2005. What is the mean of this distribution, and what is its standard deviation? Show all of your work.