THE FUNDAMENTALS OF

Political Science Research

Second Edition

Paul M. Kellstedt
Texas A&M University

Guy D. Whitten
Texas A&M University

CAMBRIDGE
UNIVERSITY PRESS

the vote nationally would get a proportionate share of the seats in the House. How many and what types of parties would you expect to see represented in the House of Representatives under this different electoral system? What theories of politics can you come up with from thinking about this hypothetical scenario?

7. *Applying formal theory to something in which you are interested.* Think about something in the political world that you would like to better understand. Try to think about the individual-level decisions that play a role in deciding the outcome of this phenomenon. What are the expected benefits and costs that the individual who is making this decision must weigh?

For exercises 8 through 11, read Robert Putnam's 1995 article "Tuning In, Tuning Out: The Strange Disappearance of Social Capital in America."

8. What is the dependent variable in Putnam's study?
9. What other possible causes of the dependent variable can you think of?
10. Can Putnam's theory be applied in other countries? Why or why not?
11. If we believe Putnam's findings, are there further implications?

3 Evaluating Causal Relationships

OVERVIEW

Modern political science fundamentally revolves around establishing whether there are *causal relationships* between important concepts. This is rarely straightforward, and serves as the basis for almost all scientific controversies. How do we know, for example, if economic development causes democratization, or if democratization causes economic development, or both, or neither? To speak more generally, if we wish to evaluate whether or not some X causes some Y, we need to cross four causal hurdles: (1) Is there a credible causal mechanism that connects X to Y? (2) Can we eliminate the possibility that Y causes X? (3) Is there covariation between X and Y? (4) Have we controlled for all confounding variables Z that might make the association between X and Y spurious? Many people, especially those in the media, make the mistake that crossing just the third causal hurdle – observing that X and Y covary – is tantamount to crossing all four. In short, finding a relationship is not the same as finding a *causal* relationship, and causality is what we care about as political scientists.

I would rather discover one causal law than be King of Persia.
– Democritus (quoted in Pearl 2000)

3.1 CAUSALITY AND EVERYDAY LANGUAGE

Like that of most sciences, the discipline of political science fundamentally revolves around evaluating causal claims. Our theories – which may be right or may be wrong – typically specify that some independent variable causes some dependent variable. We then endeavor to find appropriate empirical evidence to evaluate the degree to which this theory is or is not supported. But how do we go about evaluating causal claims? In this chapter and the next, we discuss some principles for doing this. We focus on the logic of causality and on several criteria for establishing with some confidence the degree to which a causal connection exists between two variables. Then, in Chapter 4, we discuss various ways to design research that help us to investigate causal claims. As we pursue answers to questions about causal relationships, keep our "rules of the road" from Chapter 1 in your mind, in particular the admonition to consider only empirical evidence along the way.

It is important to recognize a distinction between the nature of most scientific theories and the way the world seems to be ordered. Most of our theories are limited to descriptions of relationships between a *single* cause (the independent variable) and a *single* effect (the dependent variable). Such theories, in this sense, are very simplistic representations of reality, and necessarily so. In fact, as we noted at the end of Chapter 1, theories of this sort are laudable in one respect: They are parsimonious, the equivalent of bite-sized, digestible pieces of information. We cannot emphasize strongly enough that almost all of our theories about social and political phenomena are **bivariate** – that is, involving just two variables.

But social reality is *not* bivariate; it is **multivariate**, in the sense that any interesting dependent variable is caused by more than one factor. ("Multivariate" simply means "many variables," by which we mean involving more than two variables.) So although our theories describe the proposed relationship between some cause and some effect, we always have to keep in the forefront of our minds that the phenomenon we are trying to explain surely has many other possible causes. And when it comes time to design research to test our theoretical ideas – which is the topic of Chapter 4 – we have to try to account for, or "control for," those other causes. If we don't, then our causal inferences about whether our pet theory is right – whether X causes Y – may very well be wrong.[1] In this chapter we lay out some practical principles for evaluating whether or not, indeed, some X does cause Y. You also can apply these criteria when evaluating the causal claims made by others – be they a journalist, a candidate for office, a political scientist, a fellow classmate, a friend, or just about anyone else.

Nearly everyone, nearly every day, uses the language of causality – some of the time formally, but far more often in a very informal manner. Whenever we speak of how some event changes the course of subsequent events, we invoke causal reasoning. Even the word "because" implies that a causal process is in operation.[2] Yet, despite the ubiquitous use of the words "because," "affects," "impacts," "causes," and "causality," the meanings of these words are not exactly clear. Philosophers of science have long had vigorous debates over competing formulations of "causality."[3]

[1] Throughout this book, in the text as well as in the figures, we will use arrows as a shorthand for "causality." For example, the text "$X \rightarrow Y$" should be read as "X causes Y." Oftentimes, especially in figures, these arrows will have question marks over them, indicating that the existence of a causal connection between the concepts is uncertain.

[2] This use of terms was brought to our attention by Brady (2002).

Although our goal here is not to wade too deeply into these debates, there is one feature of the discussions about causality that deserves brief mention. Most of the philosophy of science debates originate from the world of the physical sciences. The notions of causality that come to mind in these disciplines mostly involve **deterministic relationships** – that is, relationships such that if some cause occurs, then the effect will occur *with certainty*. In contrast, though, the world of human interactions consists of **probabilistic relationships** – such that increases in X are associated with increases (or decreases) in the probability of Y occurring, but those probabilities are not certainties. Whereas physical laws like Newton's laws of motion are deterministic – think of the law of gravity here – the social sciences (including political science) more closely resemble probabilistic causation like that in Darwin's theory of natural selection, in which random mutations make an organism more or less fit to survive and reproduce.[4]

What does it mean to say that, in political science, our conceptions of causality must be probabilistic in nature? When we theorize, for example, that an individual's level of wealth causes her opinions on optimal tax policy, we do not at all mean that *every* wealthy person will want lower taxes, and *every* poor person will prefer higher taxes. Consider what would happen if we found a single rich person who favors high taxes or a single poor person who favors low taxes. (Perhaps you are, or know, such a person.) One case alone does not decrease our confidence in the theory. In this sense, the relationship is probabilistic, not deterministic. Instead of saying deterministically that "wealthy people will prefer lower taxes, and poorer people will prefer higher taxes," we say, probabilistically, that "wealthy people are more likely to prefer lower taxes, whereas poorer individuals are more likely to prefer higher taxes."

Take another example: Scholars of international conflict have noticed that there is a statistical relationship between the type of regime a country has and the likelihood of that country going to war. To be more precise, in a series of studies widely referred to as the "democratic peace" literature,

[3] You can find an excellent account of the vigor of these debates in a 2003 book by David Edmonds and John Eidinow titled *Wittgenstein's Poker: The Story of a Ten Minute Argument Between Two Great Philosophers.*

[4] Nevertheless, in reviewing three prominent attempts within the philosophy of science to elaborate on the probabilistic nature of causality, the philosopher Wesley Salmon (1993, p. 137) notes that "In the vast philosophical literature on causality [probabilistic notions of causality] are largely ignored." We borrow the helpful comparison of probabilistic social science to Darwinian natural selection from Brady (2004).

many researchers have noticed that wars are much less likely to break out between two regimes that are democracies than pairs of countries where at least one is a non-democracy. To be perfectly clear, the literature does not suggest that democracies do not engage in warfare at all, but that democracies don't fight other democracies. A variety of mechanisms have been suggested to explain this correlation, but the point here is that, if two democracies start a war with one another next year, it would be a mistake to discard the theory. A deterministic theory would say that "democracies don't go to war with one another," but a more sensible probabilistic theory would say that "democracies are highly unlikely to go to war with one another."

In political science there will always be exceptions because human beings are not deterministic robots whose behaviors always conform to lawlike statements. In other sciences in which the subjects of study do not have free will, it may make more sense to speak of laws that describe behavior. Consider the study of planetary orbits, in which scientists can precisely predict the movement of celestial bodies hundreds of years in advance. The political world, in contrast, is extremely difficult to predict. As a result, most of the time we are happy to be able to make statements about probabilistic causal relationships.

What all of this boils down to is that the entire notion of what it means for something "to cause" something else is far from a settled matter. In the face of this, should social scientists abandon the search for causal connections? Not at all. What it means is that we should proceed cautiously and with an open mind, rather than in some exceedingly rigid fashion.

3.2 FOUR HURDLES ALONG THE ROUTE TO ESTABLISHING CAUSAL RELATIONSHIPS

If we wish to investigate whether some independent variable, which we will call *X*, "causes" some dependent variable, which we will call *Y*, what procedures must we follow before we can express our degree of confidence that a causal relationship does or does not exist? Finding some sort of covariation (or, equivalently, correlation) between *X* and *Y* is not sufficient for such a conclusion.

We encourage you to bear in mind that establishing causal relationships between variables is not at all akin to hunting for DNA evidence like some episode from a television crime drama. Social reality does not lend itself to such simple, cut-and-dried answers. In light of the preceding discussion about the nature of causality itself, consider what follows to be guidelines as to what constitutes "best practice" in political science. With any theory

about a causal relationship between *X* and *Y*, we should carefully consider the answers to the following four questions:

1. Is there a credible causal mechanism that connects *X* to *Y*?
2. Can we rule out the possibility that *Y* could cause *X*?
3. Is there covariation between *X* and *Y*?
4. Have we controlled for all **confounding variables** *Z* that might make the association between *X* and *Y* **spurious**?[5]

First, we must consider whether it is believable to claim that *X could* cause *Y*. In effect, this hurdle represents an effort to answer the "how" and "why" questions about causal relationships. To do this, we need to go through a thought exercise in which we evaluate the mechanics of how *X* would cause *Y*. What is the process or mechanism that, logically speaking, suggests that *X* might be a cause of *Y*? In other words, what is it specifically about having more (or less) of *X* that will in all probability lead to more (or less) of *Y*? The more outlandish these mechanics would have to be, the less confident we are that our theory has cleared this first hurdle. Failure to clear this first hurdle is a very serious matter; the result being that either our theory needs to be thrown out altogether, or we need to revise it after some careful rethinking of the underlying mechanisms through which it works. It is worth proceeding to the second question only once we have a "yes" answer to this question.

Second, and perhaps with greater difficulty, we must ask whether we can rule out the possibility that *Y* might cause *X*. As you will learn from the discussion of the various strategies for assessing causal connections in Chapter 4, this poses thorny problems for some forms of social science research, but is less problematic for others. Occasionally, this causal hurdle can be crossed logically. For example, when considering whether a person's gender (*X*) causes him or her to have particular attitudes about abortion policy (*Y*), it is a rock-solid certainty that the reverse-causal scenario can be dismissed: A person's attitudes about abortion does not "cause" them to be male or female. If our theory does not clear this particular hurdle, the race is not lost. Under these circumstances, we should proceed to the next question, while keeping in mind the possibility that our causal arrow might be reversed.

Throughout our consideration of the first two causal hurdles, we were concerned with only two variables, *X* and *Y*. The third causal hurdle can

[5] A "confounding variable" is simply a variable that is both correlated with both the independent and dependent variable and that somehow alters the relationship between those two variables. "Spurious" means "not what it appears to be" or "false."

involve a third variable Z, and the fourth hurdle always does. Often it is the case that there are several Z variables.

For the third causal hurdle, we must consider whether X and Y covary (or, equivalently, whether they are correlated or associated). Generally speaking, for X to cause Y, there must be some form of measurable association between X and Y, such as "more of X is associated with more of Y," or "more of X is associated with less of Y." Demonstrating a simple bivariate connection between two variables is a straightforward matter, and we will cover it in Chapters 7 and 8. Of course, you may be familiar with the dictum "Correlation does not prove causality," and we wholeheartedly agree. It is worth noting, though, that correlation is normally an essential component of causality. But be careful. It is possible for a causal relationship to exist between X and Y even if there is no bivariate association between X and Y. Thus, even if we fail to clear this hurdle, we should not throw out our causal claim entirely. Instead, we should consider the possibility that there exists some confounding variable Z that we need to "control for" before we see a relationship between X and Y. Whether or not we find a bivariate relationship between X and Y, we should proceed to our fourth and final hurdle.

Fourth, in establishing causal connections between X and Y, we must face up to the reality that, as we noted at the outset of this chapter, we live in a world in which most of the interesting dependent variables are caused by more than one – often many more than one – independent variable. What problems does this pose for social science? It means that, when trying to establish whether a particular X causes a particular Y, we need to "control for" the effects of other causes of Y (and we call those other effects Z). If we fail to control for the effects of Z, we are quite likely to misunderstand the relationship between X and Y and make the wrong inference about whether X causes Y. This is the most serious mistake a social scientist can make. If we find that X and Y are correlated, but that, when we control for the effects of Z on both X and Y, the association between X and Y disappears, then the relationship between X and Y is said to be spurious.

3.2.1 Putting It All Together – Adding Up the Answers to Our Four Questions

As we have just seen, the process for evaluating a theoretical claim that X causes Y is complicated. Taken one at a time, each of the four questions in the introduction to this section can be difficult to answer with great clarity. But the challenge of evaluating a claim that X causes Y involves summing the answers to all four of these questions to determine our overall confidence about whether X causes Y. To understand this, think about the analogy that we have been using by calling these questions "hurdles." In track events that feature hurdles, runners must do their best to try to clear each hurdle as they make their way toward the finish line. Occasionally even the most experienced hurdler will knock over a hurdle. Although this slows them down and diminishes their chances of winning the race, all is not lost. If we think about putting a theory through the four hurdles posed by the preceding questions, there is no doubt our confidence will be greatest when we are able to answer all four questions the right way ("yes," "yes," "yes," "yes") and without reservation. As we described in the introduction to this section, failure to clear the first hurdle should make us stop and rethink our theory. This is also the case if we find our relationship to be spurious. For the second and third hurdles, however, failure to clear them completely does not mean that we should discard the causal claim in question. Figure 3.1 provides a summary of this process. In the subsections that follow, we will go through the process described in Figure 3.1 with a series of examples.

As we go through this process of answering the four questions, we will keep a **causal hurdles scorecard** as a shorthand for summarizing the answers to these four questions in square brackets. For now, we will limit our answers to "*y*" for "yes," "*n*" for "no," and "*?*" for "maybe." If a theory has cleared all four hurdles, the scorecard would read [*y y y y*] and the causal claim behind it would be strongly supported. As we described above, these hurdles are not all the same in terms of their impact on our assessments of causality. So, for instance, a causal claim for which the scorecard reads [*n y y y*] could be thrown out instantly. But, a claim for which it reads [*y n y y*] would have a reasonable level of evidence in its favor.

3.2.2 Identifying Causal Claims Is an Essential Thinking Skill

We want to emphasize that the logic just presented does not apply merely to political science research examples. Whenever you see a story in the news, or hear a speech by a candidate for public office, or, yes, read a research article in a political science class, it is almost always the case that some form of causal claim is embedded in the story, speech, or article. Sometimes those causal claims are explicit – indented and italicized so that you just can't miss them. Quite often, though, they are harder to spot, and most of the time not because the speaker or writer is trying to confuse you. What we want to emphasize is that spotting and identifying causal claims is a thinking skill. It does not come naturally to most people, but it can be practiced.

In our daily lives, we are often presented with causal claims by people trying to persuade us to adopt their point of view. Advocacy and attempts at persuasion, of course, are healthy features of a vibrant democracy. The health of public debate, though, will be further enhanced when citizens

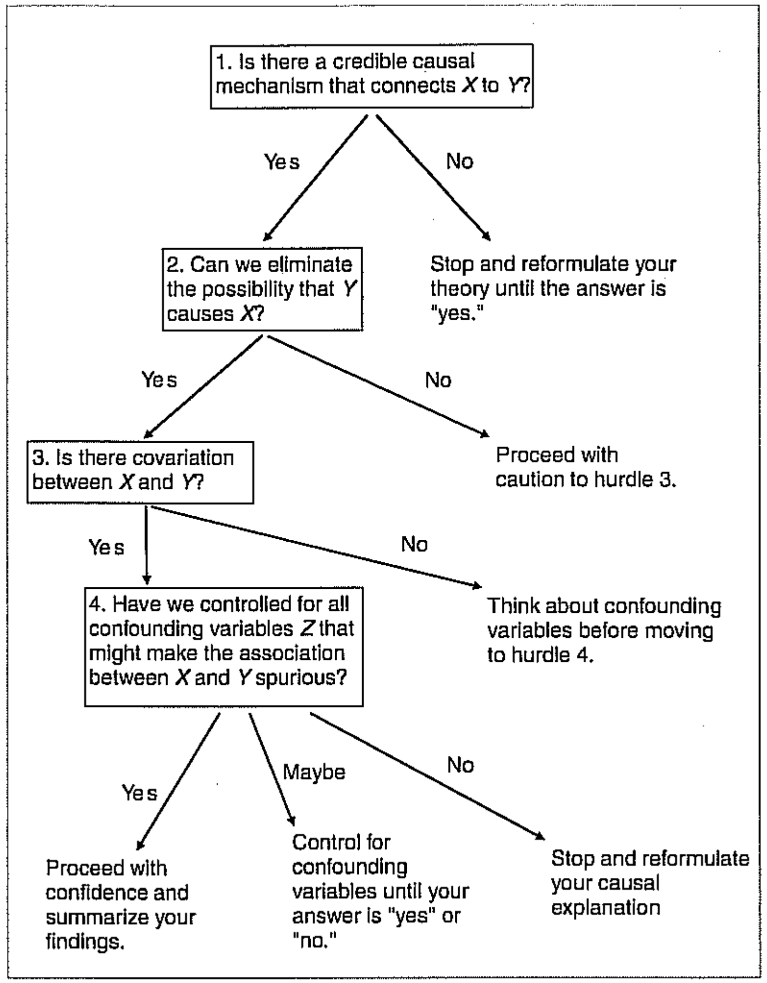

1. Is there a credible causal mechanism that connects X to Y?
Yes
No
2. Can we eliminate the possibility that Y causes X?
Stop and reformulate your theory until the answer is "yes."
Yes
No
3. Is there covariation between X and Y?
Proceed with caution to hurdle 3.
Yes
No
4. Have we controlled for all confounding variables Z that might make the association between X and Y spurious?
Think about confounding variables before moving to hurdle 4.
Yes
Maybe
No
Proceed with confidence and summarize your findings.
Control for confounding variables until your answer is "yes" or "no."
Stop and reformulate your causal explanation

Figure 3.1. The path to evaluating a causal relationship.

actively scrutinize the claims with which they are presented. Take, for example, debates in the media about the merits of private school choice programs, which have been implemented in several school districts. Among the arguments in favor of such programs is that the programs will improve student performance on standardized tests. Media reports about the successes and failures of programs like this are quite common. For example, an article in the *Washington Post* discusses a study that makes the argument that:

> African American students in the District [of Columbia] and two other cities have moved ahead of their public school classmates since they transferred to private schools with the help of vouchers, according to a new study.... The study showed that those moving to private schools scored 6 percentile points higher than those who stayed in public schools in New

> York City, Dayton, Ohio, and the District. The effect was biggest in the District, where students with vouchers moved 9 percentile points ahead of public school peers.[6]

Notice the causal claim here, which is: Participation (or not) in the school choice program (X) causes a child's test scores (Y) to vary. Often, the reader is presented with a bar chart of some sort in support of the argument. The reader is encouraged to think, sometimes subtly, that the differing heights of the bars, representing different average test scores for school choice children and public school children, means that the program *caused* the school choice children to earn higher scores. When we take such information in, we might take that nugget of evidence and be tempted to jump to the conclusion that a causal relationship exists. The key lesson here is that this is a premature conclusion.

Let's be clear: School choice programs may indeed cause students to do better on standardized tests. Our objective here is not to wade into that debate, but rather to sensitize you to the thinking skills required to evaluate the causal claim made in public by advocates such as those who support or oppose school choice programs. Evidence that students in school choice programs score higher on tests than do public school students is *one piece* of the causal puzzle – namely, it satisfies crossing hurdle three above, that there is covariation between X and Y. At this point in our evaluation, our score card reads [? ? *y* ?]. And thus, before we conclude that school choice does (or does not) cause student performance, we need to subject that claim to all four of the causal hurdles, not just the third one.

So let's apply all four causal hurdles to the question at hand. First, is there a mechanism that we can use to explain how and why attending a particular type of school – public or a voucher-sponsored private school – might affect a student's test scores? Certainly. Many private schools that participate in voucher programs have smaller class sizes (among other benefits), and smaller class sizes can translate to more learning and higher test scores. *The answer to the first question is "yes"*[*y* ? *y* ?]. Second, is it possible that the causal arrow might be reversed – that is, can we rule out the possibility that test scores cause a person to participate or not participate in a school choice program? Since the test scores occur months or even years after the person chooses a school to attend, this is not possible. *The answer to the second question is "yes"* [*y y y* ?]. Third, is there a correlation between participation in the program and test scores? The article quoted above just noted that, in the three cities considered, there is – voucher

[6] Mathews, Jay. "Scores Improve for D.C. Pupils With Vouchers" *Washington Post*, August 28, 2000, A1.

school students scored higher on standardized tests than their public school peers. *The answer to the third question is "yes"* [*y y y ?*]. Finally, have we controlled for all confounding variables that might make the association between participation in the program and test scores spurious? Remember, a potentially confounding variable is simply a variable that is related to the independent variable and is also a cause of the dependent variable. So, can we think of something that is both related to the type of school a child attends and is also a likely cause of that child's test scores? Sure. The variable "parental involvement" is a natural candidate to be a Z variable in this instance. Some children have highly involved parents – parents who read to their children, help them with homework, and take an active role in their education – while other children have parents who are much less involved. Highly involved parents are more likely than their uninvolved counterparts to learn about the existence of school choice programs in their cities, and are more likely to apply for such programs. (So Z is almost surely related to X.) And highly involved parents are more likely to create high expectations among their children, and to instill in their children a sense that achievement in school is important, all of which probably translate into having children who score better on standardized tests. (So Z is likely to be a cause of Y.) The key question then becomes: Did the study in question manage to *control for* those effects? We're a little ahead of the game here, because we haven't yet talked about the strategies that researchers employ to control for the effects of potentially confounding variables. (That task comes in Chapter 4.) But we hope you can see why controlling for the effects of parental involvement is so key in this particular situation (and in general): If our comparison of school choice children and public school children basically amounts to a comparison between the children of highly motivated parents and the children of poorly motivated parents, then it becomes very problematic to conclude that the difference between the groups' test scores was *caused by* the program. Without a control for parental involvement (Z), in other words, the relationship between school type (X) and test scores (Y) might be spurious. So, until we see evidence that this important Z has been controlled for, our scorecard for this causal claim is [*y y y n*] and we should be highly suspicious of the study's findings. More informally, without such a control, the comparison between those sets of test scores is an unfair one, because the groups would be so different in the first place. As it happens, the article from the *Washington Post* that we mentioned did include a control for parental involvement, because the students were chosen for the program by a random lottery. We'll wait until Chapter 4 to describe exactly why this makes such a big difference, but it does.

The same process can be applied to a wide variety of causal claims and questions that we encounter in our daily lives. Does drinking red wine

cause a reduction in heart disease? Does psychotherapy help people with emotional and relational problems? Do increases in government spending spur or retard economic growth? In each of these and many other examples, we might be tempted to observe a correlation between two variables and conclude that the relationship is causal. It is important for us to resist that temptation, and subject each of these claims to the more rigorous criteria that we are suggesting here. If we think about such evidence on its own in terms of our causal hurdles scorecard, what we have is [*? ? y ?*]. This is a reasonable start to the evaluation of a causal claim, but a pretty poor place to stop and draw definitive conclusions. Thinking in terms of the hurdles depicted in the scorecard, whenever someone presents us with a causal claim but fails to address each of the hurdles, we will naturally ask further questions and, when we do that, we will be much smarter consumers of information in our everyday lives.

An important part of taking a scientific approach to the study of politics is that we turn the same skeptical logic loose on scholarly claims about causal relationships. Before we can evaluate a causal theory, we need to consider how well the available evidence answers each of the four questions about X, Y, and Z. Once we have answered each of these four questions, one at a time, we then think about the overall level of confidence that we have in the claim that X causes Y.

3.2.3 What Are the Consequences of Failing to Control for Other Possible Causes?

When it comes to any causal claim, as we have just noted, the fourth causal hurdle often trips us up, and not just for evaluating political rhetoric or stories in the news media. This is true for scrutinizing scientific research as well. In fact, a substantial portion of disagreements between scholars boils down to this fourth causal hurdle. When one scholar is evaluating another's work, perhaps the most frequent objection is that the researcher "failed to control for" some potentially important cause of the dependent variable.

What happens when we fail to control for some plausible other cause of our dependent variable of interest? Quite simply, it means that we have failed to cross our fourth causal hurdle. *So long as a reasonable case can be made that some uncontrolled-for Z might be related to both X and Y, we cannot conclude with full confidence that X indeed causes Y.* Because the main goal of science is to establish whether causal connections between variables exist, then failing to control for other causes of Y is a potentially serious problem.

One of the themes of this book is that statistical analysis should not be disconnected from issues of research design – such as controlling for as many causes of the dependent variable as possible. When we discuss multiple regression (in Chapters 9, 10, and 11), which is the most common statistical technique that political scientists use in their research, the entire point of those chapters is to learn how to control for other possible causes of the dependent variable. We will see that failures of research design, such as failing to control for all relevant causes of the dependent variable, have statistical implications, and the implications are always bad. Failures of research design produce problems for statistical analysis, but hold this thought. What is important to realize for now is that good research design will make statistical analysis more credible, whereas poor research design will make it harder for any statistical analysis to be conclusive about causal connections.

3.3 WHY IS STUDYING CAUSALITY SO IMPORTANT? THREE EXAMPLES FROM POLITICAL SCIENCE

Our emphasis on causal connections should be clear. We turn now to several active controversies within the discipline of political science, showing how debates about causality lie at the heart of precisely the kinds of controversies that got you (and most of us) interested in politics in the first place.

3.3.1 Life Satisfaction and Democratic Stability

One of the enduring controversies in political science is the relationship between *life satisfaction in the mass public* and *the stability of democratic institutions*. Life satisfaction, of course, can mean many different things, but for the current discussion let us consider it as varying along a continuum, from the public's being highly unsatisfied with day-to-day life to being highly satisfied. What, if anything, is the causal connection between the two concepts?

Political scientist Ronald Inglehart (1988) argues that life satisfaction (X) *causes* democratic system stability (Y). If we think through the first of the four questions for establishing causal relationships, we can see that there is a credible causal mechanism that connects X to Y – if people in a democratic nation are more satisfied with their lives, they will be less likely to want to overthrow their government. *The answer to our first question is "yes"* [*y* ? ? ?]. Moving on to our second question: Can we eliminate the possibility that democratic stability (Y) is what causes life satisfaction (X)? We can not. It is very easy to conceive of a causal mechanism in which citizens living in stable democracies are likely to be more satisfied with their lives than citizens living in nations with a history of government instability and less-than-democratic governance. *The answer to our second question is "no"* [*y n* ? ?]. We now turn to the third question. Using an impressive amount of data from a wide variety of developed democracies, Inglehart and his colleagues have shown that there is, indeed, an association between average life satisfaction in the public and the length of uninterrupted democratic governance. That is, countries with higher average levels of life satisfaction have enjoyed longer uninterrupted periods of democratic stability. Conversely, countries with lower levels of life satisfaction have had shorter periods of democratic stability and more revolutionary upheaval. *The answer to our third question is "yes"* [*y n y* ?]. With respect to the fourth question, it is easy to imagine a myriad of other factors (Z's) that lead to democratic stability, and whether Inglehart has done an adequate job of controlling for those other factors is the subject of considerable scholarly debate. *The answer to our fourth question is "maybe"* [*y n y* ?]. Inglehart's theory has satisfactorily answered questions 1 and 3, but it is the answers to questions 2 and 4 that have given skeptics substantial reasons to doubt his causal claim.

3.3.2 Race and Political Participation in the United States

Political participation – the extent to which individual citizens engage in voluntary political activity, such as voting, working for a campaign, or making a campaign contribution – represents one of the most frequently studied facets of mass political behavior, especially in the United States. And with good reason: Participation in democratic societies is viewed by some as one measure of the health of a democracy. After decades of studying the variation in Americans' rates of participation, several demographic characteristics consistently stood out as being correlated with participation, including an individual's racial classification. Anglos, surveys consistently showed, have participated in politics considerably more frequently than either Latinos or African Americans. A comprehensive survey, for example, shows that during a typical election cycle, Anglos engaged in 2.22 "participatory acts" – such as voting, working for a campaign, making a campaign contribution, attending a protest or demonstration, and similar such activities – whereas comparable rates for African Americans and Latino citizens were 1.90 and 1.41 activities (see Verba et al. 1993, Figure 1).

Is the relationship between an individual's race (X) and the amount that the individual participates in politics (Y) a causal one? Before we accept the evidence above as conclusively demonstrating a *causal* relationship, we need to subject it to the four causal hurdles. Is there a reasonable mechanism that answers the "how" and "why" questions connecting race and political participation? There may be reason to think so. For long portions of American history, after all, some formal and many informal barriers existed prohibiting or discouraging the participation of non-Anglos. The notion that there might be residual effects of such barriers, even decades after they have been eradicated, is entirely reasonable. *The answer to our first question is "yes"* [*y* ? ? ?]. Can we eliminate the possibility that varying rates of participation cause an individual's racial classification? Obviously, yes. *The answer to our second question is "yes"* [*y y* ? ?]. Is there a correlation between an individual's race and their level of participation in the United States? The data above about the number of participatory acts among Anglos, African Americans, and Latinos clearly shows that there is a relationship; Anglos participate the most. *The answer to our third question is "yes"* [*y y y* ?]. Finally, have we controlled for all possible confounding variables *Z* that are related to both race (*X*) and participation (*Y*) that might make the relationship spurious? Verba and his colleagues suggest that there might be just such a confounding variable: socio-economic status. Less so today than in the past, socio-economic status (*Z*) is nevertheless still correlated with race (*X*). And unsurprisingly, socio-economic status (*Z*) is also a cause of political participation (*Y*); wealthy people donate more, volunteer more, and the like, than their less wealthy counterparts. Once controlling for socio-economic status, the aforementioned relationship between race and political participation entirely vanishes (see Verba et al.'s Table 8). In short, the correlation that we observe between race and political participation is spurious, or illusory; it is not a function of race, but instead a function of the disparities in wealth between Anglos and other races. Once we control for those socio-economic differences, the connection between race and participation goes away. *The answer to our fourth question is "no."* In this case, the effort to answer the fourth question actually changed our answer to the third question, moving our scorecard from [*y y y* ?] to [*y y n n*]. This is one of the important ways in which our conclusions about relationships can change when we move from a bivariate analysis in which we measure the relationship between one independent variable, *X*, and our dependent variable, *Y*, to a multiple variable analysis in which we measure the relationship between *X* and *Y* controlling for a second independent variable, *Z*. It is also possible for a lot of other things to happen when we move to controlling for *Z*. For instance, it is also possible for our scorecard to change from [*y y n n*] to [*y y y y*].

3.3.3 Evaluating Whether Head Start Is Effective

In the 1960s, as part of the War on Poverty, President Lyndon Johnson initiated the program Head Start to give economically underprivileged children a preschool experience that – the program hoped – would increase the chances that these poor children would succeed once they reached kindergarten and beyond. The program is clearly well intended, but, of course, that alone does not make it effective. Simply put: Does Head Start work? In this case, "work" would mean that Head Start could increase the chances that participants in the program would have better educational outcomes than nonparticipants.

It would be tempting, in this case, to simply compare some standardized test scores of the children who participated in Head Start with those who did not. If Head Start participants scored higher, then – voila – case closed; the program works. If not, then not. But, as before, we need to stay focused on all four causal hurdles. First, is there some credible causal mechanism that would answer the "how" and "why" questions that connect Head Start participation (*X*) to educational outcomes (*Y*)? Yes. The theory behind the program is that exposure to a preschool environment that anticipates the actual school setting helps prepare children for what they will encounter in kindergarten and beyond. Head Start, in this sense, might help reduce discipline problems, and prepare students for reading and counting, among other skills. *The answer to our first question is "yes"* [*y* ? ? ?]. Is it possible, secondly, that the causal arrow might be reversed – in other words, can we rule out the possibility that educational outcomes (*Y*) could cause participation in Head Start (*X*)? Because testing would take place years after participation in the program, yes. *The answer to our second question is "yes"* [*y y* ? ?]. Is there an association between participation in the program and learning outcomes? Study after study has shown that Head Start participants fare better when tested, and have fewer instances of repeating a grade, than those who have no preschool experience. For example, a widely cited study shows that Head Start children do better on a vocabulary test suitable for young children than do students who have no preschool experience (Currie and Thomas 1995). *The answer to our third question is "yes"* [*y y y* ?]. But, as was the case with the school-voucher example discussed previously, a potentially confounding variable – parental involvement (*Z*) – lurks nearby. Highly involved parents (*Z*) are more likely to seek out, be aware of, and enroll their children (*X*) in programs like Head Start that might benefit their children. Parents who are less involved in their childrens' lives are less likely to avail themselves of the potential opportunities that Head Start creates. And, as before, highly involved parents (*Z*) are likely to have positive effects on their children's educational outcomes. The key question, then, becomes: Do parental effects (*Z*) make the relationship between Head Start and later educational outcomes spurious? The aforementioned study by Currie and Thomas uses both statistical controls as well as controls in the design of their research to account for parental factors, and they find that Head Start has lasting educational effects only for Anglo children, but not for African American children (see their Table 4). Again, that phrase "statistical controls" may not be quite as transparent as it will be later on in this book. For now, suffice it to say that these researchers used all of the techniques available to them to show that Head Start does, indeed, have positive effects for some, but not all, children. *The answer to our fourth question is a highly qualified "yes"* [*y y y y*].

3.4 WRAPPING UP

Learning the thinking skills required to evaluate causal claims as conclusively as possible requires practice. They are intellectual habits that, like a good knife, will sharpen with use.

Translating these thinking skills into actively designing new research that helps to address causal questions is the subject of Chapter 4. All of the "research designs" that you will learn in that chapter are strongly linked to issues of evaluating causal claims. Keeping the lessons of this chapter in mind as we move forward is essential to making you a better consumer of information, as well as edging you forward toward being a producer of research.

CONCEPTS INTRODUCED IN THIS CHAPTER

- bivariate – involving just two variables.
- causal hurdles scorecard – a shorthand for summarizing evidence about whether an independent variable causes a dependent variable.
- confounding variable – a variable that is correlated with both the independent and dependent variables and that somehow alters the relationship between those two variables.
- deterministic relationship – if some cause occurs, then the effect will occur with certainty.
- multivariate – involving more than two variables.
- probabilistic relationship – increases in *X* are associated with increases (or decreases) in the probability of *Y* occurring, but those probabilities are not certainties.
- spurious – not what it appears to be, or false.

EXERCISES

1. Think back to a history class in which you learned about the "causes" of a particular historical event (for instance, the Great Depression, the French Revolution, or World War I). How well does each causal claim perform when you try to answer the four questions for establishing causal relationships?

2. Go to your local newspaper's web site (if it has one; if not, pick the web site of any media outlet you visit frequently). In the site's "Search" box, type the words "research cause" (without quotes). (*Hint:* You may need to limit the search time frame, depending on the site you visit.) From the search results, find two articles that make claims about causal relationships. Print them out, and include a brief synopsis of the causal claim embedded in the article.

3. For each of the following examples, imagine that some researcher has found the reported pattern of covariation between *X* and *Y*. Can you think of a variable *Z* that might make the relationship between *X* and *Y* spurious?

(a) The more firefighters (*X*) that go to a house fire, the greater property damage that occurs (*Y*).

(b) The more money spent by an incumbent member of Congress's campaign (*X*), the lower their percentage of vote (*Y*).

(c) Increased consumption of coffee (*X*) reduces the risk of depression among women (*Y*).

(d) The higher the salaries of Presbyterian ministers (*X*), the higher the price of rum in Havana (*Y*).

4. For each of the following pairs of independent and dependent variables, write about both a probabilistic and a deterministic relationship to describe the likely relationship:

(a) A person's education (*X*) and voter turnout (*Y*).

(b) A nation's economic health (*X*) and political revolution (*Y*).

(c) Candidate height (*X*) and election outcome (*Y*).

5. Take a look at the codebook for the data set "BES 2005 Subset" and write about your answers to the following items:

(a) Develop a causal theory about the relationship between an independent variable (*X*) and a dependent variable (*Y*) from this data set. Is it the credible causal mechanism that connects *X* to *Y*? Explain your answer.

(b) Could *Y* cause *X*? Explain your answer.

(c) What other variables (*Z*) would you like to control for in your tests of this theory?

6. Imagine causal claims for which the scorecards are listed below. Which of these clams would you evaluate as most strongly supported? Explain your answer.

(a) [*y n y y*]

(b) [*y y y n*]

(c) [*? y y y*]

7. Researcher A and Researcher B are having a scientific debate. What are they arguing about if their argument is focused on:

(a) causal hurdle 1

(b) causal hurdle 2

(c) causal hurdle 3

(d) causal hurdle 4

8. Find a political science journal article of interest to you, and of which your instructor approves, and answer the following items (be sure to provide a full citation to the chosen article with your answers):
   (a) Briefly describe the causal theory that connects the independent and dependent variables.
   (b) Create a causal hurdles scorecard for this theory and write an explanation for each of your entries in the scorecard.

4 Research Design

OVERVIEW

Given our focus on causality, what research strategies do political scientists use to investigate causal relationships? Generally speaking, the controlled experiment is the foundation for scientific research. And some political scientists use experiments in their work. However, owing to the nature of our subject matter, most political scientists adopt one of two types of "observational" research designs that are intended to mimic experiments. The cross-sectional observational study focuses on variation across individual units (like people or countries). The time-series observational study focuses on variation in aggregate quantities (like presidential popularity) over time. What is an "experiment" and why is it so useful? How do observational studies try to mimic experimental designs? Most importantly, what are the strengths and weaknesses of each of these three research designs in establishing whether or not causal relationships exist between concepts? That is, how does each one help us to get across the four causal hurdles identified in Chapter 3? Relatedly, we introduce issues concerning the selection of samples of cases to study in which we are not able to study the entire population of cases to which our theory applies. This is a subject that will feature prominently in many of the subsequent chapters.

4.1 COMPARISON AS THE KEY TO ESTABLISHING CAUSAL RELATIONSHIPS

So far, you have learned that political scientists care about causal relationships. You have learned that most phenomena we are interested in explaining have multiple causes, but our theories typically deal with only one of them while ignoring the others. In some of the research examples in the previous chapters, we have noted that the multivariate nature of the world can make our first glances at evidence misleading. In the example dealing with race and political participation, at first it appeared that race might be causally related to participation rates, with Anglos participating more than those of other races. But, we argued, in this particular case, the first glance was potentially quite misleading.

Why? Because what appeared to be the straightforward comparisons between three groups – participation rates between Anglos, Latinos, and African Americans – ended up being far from simple. On some very important factors, our different groupings for our independent variable X were far from equal. That is, people of different racial groupings (X) had differing socio-economic statuses (Z), which are correlated with race (X) and also affected their levels of participation (Y). As convincing as those bivariate comparisons might have been, they would likely be misleading.

Comparisons are at the heart of science. If we are evaluating a theory about the relationship between some X and some Y, the scientist's job is to do everything possible to make sure that no other influences (Z) interfere with the comparisons that we will rely on to make our inferences about a possible causal relationship between X and Y.

The obstacles to causal inference that we described in Chapter 3 are substantial, but surmountable. We don't know whether, in reality, X causes Y. We may be armed with a theory that suggests that X does, indeed, cause Y, but theories can be (and often are) wrong or incomplete. So how do scientists generally, and political scientists in particular, go about testing whether X causes Y? There are several strategies, or **research designs**, that researchers can use toward that end. The goal of all types of research designs is to help us evaluate how well a theory fares as it makes its way over the four causal hurdles – that is, to answer as conclusively as is possible the question about whether X causes Y. In the next two sections we focus on the two strategies that political scientists use most commonly and effectively: **experiments** and **observational studies**.[1]

4.2 EXPERIMENTAL RESEARCH DESIGNS

Suppose that you were a candidate for political office locked in what seems to be a tight race. Your campaign budget has money for the end of the campaign, and you're deciding whether or not to make some television ad buys for a spot that sharply contrast your record with your opponent's – what some will surely call a negative, attack ad. The campaign manager has had a public relations firm craft the spot, and has shown it to you in your strategy meetings. You like it, but you look to your staff and ask the bottom-line question: "Will the ad work with the voters?" In effect, you have two choices: run the attack ad, or do nothing.

[1] Throughout this book, we will use the term "experiment" in the same way that researchers in medical science use the term "randomized control trial."

We hope that you're becoming accustomed to spotting the causal questions embedded in this scenario: Exposure to a candidate's negative ad (X) may, or may not, affect a voter's likelihood of voting for that candidate (Y). And it is important to add here that the causal claim has a particular directional component to it; that is, exposure to the advertisement will *increase* the chances that a voter will choose that candidate.[2]

How might researchers in the social sciences evaluate such a causal claim? Those of you who are campaign junkies are probably thinking that your campaign would run a focus group to see how some voters react to the ad. And that's not a bad idea. Let's informally define a focus group as a group of subjects selected to expose to some idea (like a new kitchen knife or a candidate's TV ad), and to try to gather the subjects' responses to the idea. There's a problem with the focus group, though, particularly in the case at hand of the candidate's TV ad: What would the subjects have said about the candidate had they *not* been exposed to the ad? There's nothing to use as a basis for comparison.

It is very important, and not at all surprising, to realize that voters may vote either for or against you for a variety of reasons (Z's) that have nothing to do with exposure to the advertisements – varying socio-economic statuses, varying ideologies, and party identifications can all cause voters to favor one candidate over another. So how can we establish whether, among these other influences (Z), the advertisement (X) also causes voters to be more likely to vote for you (Y)?

Can we do better than the focus group? What would a more scientific approach look like? As the introduction to this chapter highlights, we will need a comparison of some kind, and we will want that comparison to isolate any potentially different effects that the ad has on a person's likelihood of voting for you.

The standard approach to a situation like this in the physical and medical sciences is that we would need to conduct an experiment. Because the word "experiment" has such common usage, its scientific meaning is frequently misunderstood. An experiment is *not* simply any kind of analysis that is quantitative in nature; neither is it exclusively the domain of laboratories and white-coated scientists with pocket protectors. We define an

[2] There is a substantial literature in political science about the effects that negative advertisements have on both voter turnout and vote choice. For contrasting views on the effects of negative ads, see Ansolabehere and Iyengar (1997), Wattenberg and Brian (1999), and Geer (2006).

experiment as follows: *An experiment is a research design in which the researcher both controls and randomly assigns values of the independent variable to the participants.*

Notice the twin components of the definition of the experiment: that the researcher both *controls* values of the independent variable – or X, as we have called it – as well as *randomly assigns* those values to the participants in the experiment. Together, these two features form a complete definition of an experiment, which means that there are no other essential features of an experiment beside these two.

What does it mean to say that a researcher "controls" the value of the independent variable that the participants receive? It means, most importantly, that the values of the independent variable that the participants receive are *not* determined either by the participants themselves or by nature. In our example of the campaign's TV ad, this requirement means that we cannot compare people who, by their own choice, already have chosen to expose themselves to the TV ad (perhaps because they're political junkies and watch a lot of cable news programs, where such ads are likely to air). It means that we, the researchers, have to decide which of our experimental participants will see the ads and which ones will not.

But the definition of an experiment has one other essential component as well: We, the researchers, must not only control the values of the independent variable, but *we must also assign those values to participants randomly*. In the context of our campaign ad example, this means that we must toss coins, draw numbers out of a hat, use a random-number generator, or some other such mechanism to divide our participants into a **treatment group** (who will see the negative ad) and a **control group** (who will not see the ad, but will instead watch something innocuous, in a social science parallel to a **placebo**).

What's the big deal here? Why is randomly assigning subjects to treatment groups important? What scientific benefits arise from the random assignment of people to treatment groups? To see why this is so crucial, recall that we have emphasized that all science is about comparisons and also that every interesting phenomenon worth exploring – every interesting dependent variable – is caused by many factors, not just one. Random assignment to treatment groups ensures that the comparison we make between the treatment group and the control group is as pure as possible and that some other cause (Z) of the dependent variable will not pollute that comparison. By first taking a group of participants and then randomly splitting them into two groups on the basis of a coin flip, what we have ensured is that the participants will not be systematically different from one another. Indeed, provided that the participant pool is reasonably large, randomly assigning participants to treatment groups ensures that the groups, as a whole, are *identical*. If the two groups are identical, save for the coin flip, then we can be certain that any differences we observe in the groups must be because of the independent variable that we have assigned to them.

Return to our campaign advertising example. An experiment involving our new ad would involve finding a group of people – however obtained – and then randomly assigning them to view either our new ad or something that is not related to the campaign (like a cartoon or a public service announcement). We fully realize that there are other causes of people's voting behaviors and that our experiment does not negate those factors. In fact, our experiment will have nothing whatsoever to say about those other causes. What it *will* do, and do well, is to determine whether our advertisement had a positive or negative effect, or none at all, on voter preferences.

Contrast the comparison that results from an experiment with a comparison that arises from a non-experiment. (We'll discuss non-experimental designs in the next section.) Suppose that we don't do an experiment and just run the ad, and then spend our campaign money conducting a survey asking people if they've seen our ad, and for whom they plan to vote. Let's even assume that, in conducting our survey, we obtain a random sample of citizens in the district where the election will take place. If we analyze the results of the survey and discover that, as hoped, the people who say that they have seen our ad are more likely to vote for us than people who say they have not seen our ad, does that mean that the ad *caused* – see that word again? – people's opinions to shift in our favor? No. Why not? Because people who saw our ad and people who did not see our ad might be *systematically different* from one another. What does that mean? It means that people who voluntarily watch a lot of politics on TV are (of course) more interested in politics than those who watch the rest of what appears on TV. In this case, a person's level of interest in politics could be an important Z variable. Interest in politics could very well be associated with a person's likelihood to vote for you. What this means is that the simple comparison in a non-experiment between those who do and do not see the ad is potentially misleading because it is confounded by other factors like interest in politics. So is the higher support for you the result of the advertisement, or is it the result of the fact that people likely to see the ad in the first place are people with higher interest in politics? Because this particular non-experimental research design does not answer that question, it does not clear our fourth causal hurdle. It is impossible to know whether it was the ad that caused the voters to support you. In this non-experimental design just described, because there are other factors that influence support for a candidate – and, critically, because these factors are also related to whether or not people will see the advertisement – it is very difficult to say conclusively that

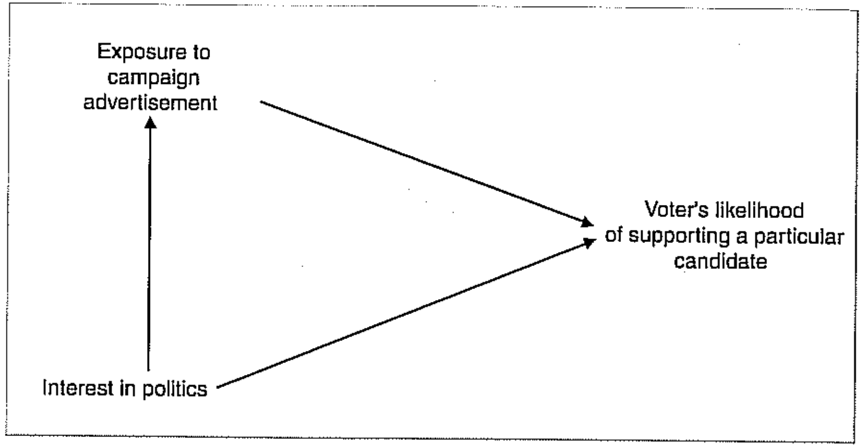

Exposure to campaign advertisement
Voter's likelihood of supporting a particular candidate
Interest in politics

Figure 4.1. The possibly confounding effects of political interest in the advertisement viewing–vote intention relationship.

the independent variable (ad exposure) causes the dependent variable (vote intention). Figure 4.1 shows this graphically.

Here is where experiments differ so drastically from any other kind of research design. What experimental research designs accomplish by way of random assignment to treatment groups, then, is to decontaminate the comparison between the treatment and control group of all other influences. Before any stimulus (like a treatment or placebo) is administered, all of the participants are in the same pool. Researchers divide them by using some random factor like a coin flip, and that difference is the only difference between the two groups.

Think of it another way. The way that the confounding variables in Figure 4.1 are correlated with the independent variable is highly improbable in an experiment. Why? Because if X is determined by randomness, like a coin flip, then (by the very definition of randomness) it is exceedingly unlikely to be correlated with anything (including confounding variables Z). When researchers control and assign values of X randomly, the comparison between the different groups will not be affected by the fact that other factors certainly do cause Y, the dependent variable. In an experiment, then, because X is only caused by randomness, it means that we can erase the connection between Z and X in Figure 4.1. And, recalling our definition of a confounding variable, if Z is not correlated with X, it cannot confound the relationship between X and Y.

Connect this back to our discussion from Chapter 3 about how researchers attempt to cross four hurdles in their efforts to establish whether some X causes Y. As we will see, experiments are not the only method that help researchers cross the four causal hurdles, but they are uniquely capable in accomplishing important parts of that task. Consider each hurdle in turn. First, we should evaluate whether there is a credible causal mechanism before we decide to run the experiment. It is worth noting that the crossing of this causal hurdle is neither easier nor harder in experiments than in non-experiments. Coming up with a credible causal scenario that links X to Y heightens our dependence on theory, not on data or research design.

Second, in an experiment, it is impossible for Y to cause X – the second causal hurdle – for two reasons. First, assigning X occurs in time before Y is measured, which makes it impossible for Y to cause X. More importantly, though, as previously noted, if X is generated by randomness alone, then nothing (including Y) can cause it. So, in Figure 4.1, we could eliminate any possible reverse-causal arrow flowing from Y to X.

Establishing, third, whether X and Y are correlated is similarly easy regardless of chosen research design, experimental or non-experimental (as we will see in Chapter 7). What about our fourth causal hurdle? Have we controlled for all confounding variables Z that might make the association between X and Y spurious? Experiments are uniquely well equipped to help us answer this question definitively. An experiment does not, in any way, eliminate the possibility that a variety of other variables (that we call Z) might also affect Y (as well as X). What the experiment does, through the process of randomly assigning subjects to different values of X, is to equate the treatment and control groups on all possible factors. On every possible variable, whether or not it is related to X, or to Y, or to both, or to neither, the treatment and control groups should, in theory, be identical. That makes the comparison between the two values of X unpolluted by any possible Z variables because we expect the groups to be equivalent on all values of Z.

Remarkably, the experimental ability to control for the effects of outside variables (Z) applies to *all* possible confounding variables, regardless of whether we, the researchers, are aware of them. Let's make the example downright preposterous. Let's say that, 20 years from now, another team of scientists discovers that having attached (as opposed to detached) earlobes causes people to have different voting behaviors. Does that possibility threaten the inference that we draw from our experiment about our campaign ad? No, not at all. Why not? Because, whether or not we are aware of it, the random assignment of participants to treatment groups means that, whether we are paying attention to it or not, we would expect our treatment and control groups to have equal numbers of people with attached earlobes, and for both groups to have equal numbers of people with detached earlobes. The key element of an experimental research design – randomly assigning subjects to different values of X, the independent variable – controls for every Z in the universe, whether or not we are aware of that Z.

In summary, if we think back to the causal hurdles scorecard from the previous chapter, all properly set-up experiments start out with a scorecard reading [? y ? y]. The ability of experimental designs to cleanly and definitively answer "yes" to the fourth hurdle question – Have we controlled for all confounding variables Z that might make the association between X and Y spurious? – is a massive advantage.[3] All that remains for establishing a causal relationship is the answers to clear the first hurdle – Is there a credible causal mechanism that connects X to Y? – and hurdle three – Is there covariation between X and Y? The difficulty of clearing hurdle one is unchanged, but the third hurdle is much easier because we need only to make a statistical evaluation of the relationship between X and Y. As we will see in Chapter 7, such evaluations are pretty straightforward, especially when compared to statistical tests that involve controlling for other variables (Z).

Together, all of this means that experiments bring with them a particularly strong confidence in the causal inferences drawn from the analysis. In scientific parlance, this is called **internal validity**. If a research design produces high levels of confidence in the conclusions about causality, it is said to have high internal validity. Conversely, research designs that do not allow for particularly definitive conclusions about whether X causes Y are said to have low degrees of internal validity.

4.2.1 "Random Assignment" versus "Random Sampling"

It is critical that you do not confuse *the experimental process of randomly assigning subjects to treatment groups*, on the one hand, with *the process of randomly sampling subjects for participation*, on the other hand. They are entirely different, and in fact have nothing more in common than that six-letter word "random." They are, however, quite often confused for one another. **Random assignment** to treatment and control groups occurs when the participants for an experiment are assigned randomly to one of several possible values of X, the independent variable. Importantly, this definition says nothing at all about how the subjects were selected for participation. But **random sampling** is, at its very heart, about how researchers select cases for inclusion in a study – they are selected at random, which means that every member of the underlying **population** has an equal probability of being selected. (This is common in survey research, for example.)

[3] After all, even the best designed and executed non-experimental designs must remain open to the possibility that, somewhere out there, there is a Z variable that has not yet been considered and controlled for.

Mixing up these two critical concepts will produce a good bit of confusion. In particular, confusing random sampling with random assignment to treatment groups will mean that the distinction between experiments and non-experiments has been lost, and this difference is among the more important ones in all of science. To understand how science works, keep these two very important concepts separate from one another.

4.2.2 Varieties of Experiments and Near-Experiments

Not all experiments take place in a laboratory with scientists wearing white lab coats. Some experiments in the social sciences are conducted by surveys that do use random samples (see above). Since 1990 or so, there has been a growing movement in the field of survey research – which has traditionally used random samples of the population – to use computers in the interviewing process that includes experimental randomization of variations in survey questions, in a technique called a **survey experiment**. Such designs are intended to reap the benefits of both random assignment to treatment groups, and hence have high internal validity, as well as the benefits of a random sample, and hence have high **external validity**.[4] Survey experiments may be conducted over the phone or, increasingly, over the internet.

Another setting for an experiment is out in the natural world. A **field experiment** is one that occurs in the natural setting where the subjects normally lead their lives. Random assignment to treatment groups has enabled researchers in the social sciences to study subjects that seemed beyond the reach of experimentation. Economists have long sought conclusive evidence about the effectiveness (or the lack thereof) of economic development policies. For example, do government fertilizer subsidies (X) affect agricultural output (Y)? Duflo, Kremer, and Robinson (2011) report the results of an experiment in a region in Western Kenya in which a subsidy of free delivery of fertilizer was offered only to randomly chosen farmers, but not to others.

Field experiments can also take place in public policy settings, sometimes with understandable controversy. Does the police officer's decision whether or not to arrest the male at a domestic violence call (X) affect the incidence of repeat violence at the same address in the subsequent months (Y)? Sherman and Berk (1984) conducted a field experiment in Minneapolis, randomizing whether or not the male in the household would automatically (or not) be arrested when police arrived at the house.

On occasion, situations in nature that are not properly defined as experiments – because the values of X have not been controlled and assigned

[4] See Piazza, Sniderman, and Tetlock (1990) and Sniderman and Piazza (1993).

by the researcher – nevertheless resemble experiments in key ways. In a **natural experiment** – which, we emphasize, does not meet our definition of an experiment – values of the independent variable arise naturally in such a way as to make it seem as if true random assignment by a researcher has occurred. For example, does the size of an ethnic group within a population (X) affect inter-group conflict or cooperation (Y)? Posner (2004) investigates why the Chewa and Tumbuka peoples are allies in Zambia but are adversaries in Malawi. Because the sizes of the groups in the different countries seem to have arisen randomly, the comparison is treated *as if* the sizes of the respective populations were assigned randomly by the researcher, when (of course) they were not.

4.2.3 Are There Drawbacks to Experimental Research Designs?

Experiments, as we have seen, have a unique ability to get social scientists across our hurdles needed to establish whether X causes Y. But that does not mean they are without disadvantages. Many of these disadvantages are related to the differences between medical and physical sciences, on the one hand, and the social sciences, on the other. We now discuss four drawbacks to experimentation.

First, especially in the social sciences, not every independent variable (X) is controllable and subject to experimental manipulation. Suppose, for example, that we wish to study the effects of gender on political participation. Do men contribute more money, vote more, volunteer more in campaigns, than women? There are a variety of non-experimental ways to study this relationship, but it is impossible to experimentally manipulate a subject's gender. Recall that the definition of an experiment is that the researcher both controls and randomly assigns the values of the independent variable. In this case, the presumed cause (the independent variable) is a person's gender. Compared with drugs versus placebos, assigning a participant's gender is another matter entirely. It is, to put it mildly, impossible. People show up at an experiment either male or female, and it is not within the experimenter's power to "randomly assign" a participant to be male or female.

This is true in many, many political science examples. There are simply a myriad of substantive problems that are impossible to study in an experimental fashion. How does a person's partisanship (X) affect his issue opinions (Y)? How does a person's income level (X) affect her campaign contributions (Y)? How does a country's level of democratization (X) affect its openness to international trade (Y)? How does the level of military spending in India (X) affect the level of military spending in Pakistan (Y) – and, for that matter, vice versa? How does media coverage (X) in an election campaign influence voters' priorities (Y)? Does serving in the UK parliament (X) make members of parliament wealthy (Y)? In each of these examples that intrigues social scientists, the independent variable is simply not subject to experimental manipulation. Social scientists cannot, in any meaningful sense, "assign" people a party identification or an income, "assign" a country a level of democratization or level of military spending, "assign" a campaign-specific, long-term amount of media coverage, or "assign" different candidates to win seats in parliament. These variables simply exist in nature, and we cannot control exposure to them and randomly assign different values to different cases (that is, individual people or countries). And yet, social scientists feel compelled to study these phenomena, which means that, in those circumstances, we must turn to a non-experimental research design.

A second potential disadvantage of experimental research designs is that experiments often suffer from low degrees of external validity. We have noted that the key strength of experiments is that they typically have high levels of internal validity. That is, we can be quite confident that the conclusions about causality reached in the analysis are not confounded by other variables. External validity, in a sense, is the other side of the coin, as it represents the degree to which we can be confident that the results of our analysis apply not only to the participants in the study, but also to the population more broadly construed.

There are actually two types of concerns with respect to external validity. The first is the external validity of the sample itself. Recall that there is nothing whatsoever in our definition of an experiment that describes how researchers recruit or select people to participate in the experiment. To reiterate: *It is absolutely not the case that experiments require a random sample of the target population.* Indeed, it is extremely rare for experiments to draw a random sample from a population. In drug-trial experiments, for example, it is common to place advertisements in newspapers or on the radio to invite participation, usually involving some form of compensation to the participants. Clearly, people who see and respond to advertisements like this are not a random sample of the population of interest, which is typically thought of as all potential recipients of the drug. Similarly, when professors "recruit" people from their (or their colleagues') classes, the participants are not a random sample of *any* population.[5] The participant pool

[5] Think about that for a moment. Experiments in undergraduate psychology or political science classes are not a random sample of 18- to 22-year-olds, or even a random sample of undergraduate students, or even a random sample of students from your college or university. Your psychology class is populated with people more interested in the social sciences than in the physical sciences or engineering or the humanities.

in this case represents what we would call a **sample of convenience**, which is to say, this is more or less the group of people we could beg, coerce, entice, or cajole to participate.

With a sample of convenience, it is simply unclear how, if at all, the results of the experiment generalize to a broader population. As we will learn in Chapter 6, this is a critical issue in the social sciences. Because most experiments make use of such samples of convenience, with any single experiment, it is difficult to know whether the results of that analysis are in any way typical of what we would find in a different sample. With experimental designs, then, scientists learn about how their results apply to a broader population through the process of **replication**, in which researchers implement the same procedures repeatedly in identical form to see if the relationships hold in a consistent fashion.

There is a second external validity concern with experiments that is more subtle, but perhaps just as important. It concerns the external validity of the stimulus. To continue our example of whether the campaign ad affects voter intentions, if we were to run an experiment to address this question, what would we do? First, we would need to obtain a sample of volunteer subjects somehow. (Remember, they need not be a random sample.) Second, we would divide them, on a random basis, into experimental and control groups. We would then sit them in a lab in front of computers, and show the ad to the experimental group, and show something innocuous to the control group. Then we would ask the subjects from both groups their vote intentions, and make a comparison between our groups. Just as we might have concerns about how externally valid our sample is, because they may not be representative of the underlying population, we should also be concerned about how externally valid our stimulus is. What do we mean here? The stimulus is the X variable. In this case, it is the act of sitting the experimental and control subjects down and having them watch (different) video messages on the computer screens. How similar is that stimulus to one that a person experiences in his or her home – that is, in their more natural environment? In some respects it is quite different. In our hypothetical experiment, the individual does not choose what he or she sees. The exposure to the ad is forced (once the subject consents to participate in the experiment). At home? People who don't want to be exposed to political ads can avoid them rather easily if they so choose, simply by not watching particular channels or programs, or by not watching TV at all, or by flipping the channel when a political ad starts up. But the comparison in our hypothetical experiment is entirely insensitive to this key difference between the experimental environment and the subject's more natural environment. To the extent that an experiment creates an entirely artificial environment, we might be concerned that the results of that experiment will be found in a more real-world context.[6]

Experimental research designs, at times, can be plagued with a third disadvantage, namely that they carry special ethical dilemmas for the researcher. Ethical issues about the treatment of human participants occur frequently with medical experiments, of course. If we wished to study experimentally the effects of different types of cancer treatments on survival rates, this would require obtaining a sample of patients with cancer and then randomly assigning the patients to differing treatment regimens. This is typically not considered acceptable medical practice. In such high-stakes medical situations, most individuals value making these decisions themselves, in consultation with their doctor, and would not relinquish the important decisions about their treatment to a random-number generator.

Ethical situations arise less frequently, and typically less dramatically, in social science experimentation, but they do arise on occasion. During the behavioral revolution in psychology in the 1960s, several famous experiments conducted at universities produced vigorous ethical debates. Psychologist Stanley Milgram (1974) conducted experiments on how easily he could make individuals obey an authority figure. In this case, the dependent variable was the willingness of the participant to administer what he or she believed to be a shock to another participant, who was in fact an employee of Milgram's. (The ruse was that Milgram told the participant that he was testing how negative reinforcement – electric shocks – affected the "learning" of the "student.") The independent variable was the degree to which Milgram conveyed his status as an authority figure. In other words, the X that Milgram manipulated was the degree to which he presented himself as an authority who must be obeyed. For some participants, Milgram wore a white lab coat and informed them that he was a professor at Yale University. For others, he dressed more casually and never mentioned his institutional affiliation. The dependent variable, then, was how strong the (fake) shocks would be before the subject simply refused to go on. At the highest extreme, the instrument that delivered the "shock" said "450 volts, XXX." The results of the experiment were fascinating because, to his surprise, Milgram found that the great majority of his participants were willing to administer even these extreme shocks to the "learners." But scientific review boards consider such experiments unethical today, because

[6] For a discussion of the external validity of experiments embedded in national surveys, see Barabas and Jerit (2010). For a substantive application where the issues of external validity of the stimulus are pivotal in determining the results of the experiment, see Arceneaux and Johnson (2011). See also Morton and Williams (2010, p. 264), who refer to this problem as one of "ecological validity."

the experiment created a great degree of emotional distress among the true participants.

A fourth potential drawback of experimental research designs is that, when interpreting the results of an experiment, we sometimes make mistakes of emphasis. If an experiment produces a finding that some *X* does indeed cause *Y*, that does not mean that that particular *X* is the most prominent cause of *Y*. As we have emphasized repeatedly, a variety of independent variables are causally related to every interesting dependent variable in the social sciences. Experimental research designs often do not help to sort out which causes of the dependent variable have the largest effects and which ones have smaller effects.

4.3 OBSERVATIONAL STUDIES (IN TWO FLAVORS)

Taken together, the drawbacks of experiments mean that, for any given political science research situation, implementing an experiment often proves to be unworkable, and sometimes downright impossible. As a result, experimentation is not the most common research design used by political scientists. In some subfields, such as political psychology – which, as the name implies, studies the cognitive and emotional underpinnings of political decision making – experimentation is quite common. And it is becoming more common in the study of public opinion and electoral competition. But the experiment, for many researchers and for varying reasons, remains a tool that is not applicable to many of the phenomena that we seek to study.

Does this mean that researchers have to shrug their shoulders and abandon their search for causal connections before they even begin? Not at all. But what options do scholars have when they cannot control exposure to different values of the independent variables? In such cases, the only choice is to take the world as it already exists and make the comparison between either individual units – like people, political parties, or countries – or between an **aggregate** quantity that varies over time. These represent two variants of what is most commonly called an observational study. Observational studies are not experiments, but they seek to emulate them. They are known as observational studies because, unlike the controlled and somewhat artificial nature of most experiments, in these research designs, researchers simply take reality as it is and "observe" it, attempting to sort out causal connections without the benefit of randomly assigning participants to treatment groups. Instead, different values of the independent variable already exist in the world, and what scientists do is observe them and then evaluate their theoretical claims by putting them through the same four causal hurdles to discover whether *X* causes *Y*.

This leads to the definition of an observational study: An observational study is a research design in which the researcher does *not* have control over values of the independent variable, which occur naturally. However, it is necessary that there be some degree of variability in the independent variable across cases, as well as variation in the dependent variable.

Because there is no random assignment to treatment groups, as in experiments, some scholars claim that it is impossible to speak of causality in observational studies, and therefore sometimes refer to them as **correlational studies**. Along with most political scientists, we do not share this view. Certainly experiments produce higher degrees of confidence about causal matters than do observational studies. However, in observational studies, if sufficient attention is paid to accounting for all of the other possible causes of the dependent variable that are suggested by current understanding, then we can make informed evaluations of our confidence that the independent variable does cause the dependent variable.

Observational studies, as this discussion implies, face exactly the same four causal hurdles as do experiments. (Recall that those hurdles are present in any research design.) So how, in observational studies, do we cross these hurdles? The first causal hurdle – Is there a credible mechanism connecting *X* and *Y*? – is identical in experimental and observational studies.

In an observational study, however, crossing the second causal hurdle – Can we eliminate the possibility that *Y* causes *X*? – can sometimes be problematic. For example, do countries with higher levels of economic development (*X*) have, as a consequence, more stable democratic regimes (*Y*)? Crossing the second causal hurdle, in this case, is a rather dicey matter. It is clearly plausible that having a stable democratic government makes economic prosperity more likely, which is the reverse-causal scenario. After all, investors are probably more comfortable taking risks with their money in democratic regimes than in autocratic ones. Those risks, in turn, likely produce greater degrees of economic prosperity. It is possible, of course, that *X* and *Y* are mutually reinforcing – that is, *X* causes *Y* and *Y* causes *X*.

The third hurdle – Is there covariation between *X* and *Y*? – is, as we mentioned, no more difficult for an observational study than for an experiment. (The techniques for examining relationships between two variables are straightforward, and you will learn them in Chapters 7 and 8.) But, unlike in an experimental setting, if we fail to find covariation between *X* and *Y* in an observational setting, we should still proceed to the fourth hurdle because the possibility remains that we will find covariation between and *X* and *Y* once we control for some variable *Z*.

The most pointed comparison between experiments and observational studies, though, occurs with respect to the fourth causal hurdle. The near-magic that happens in experiments because of random assignment to treatment groups – which enables researchers to know that no other factors interfere in the relationship between X and Y – is not present in an observational study. So, in an observational study, the comparison between groups with different values of the independent variable may very well be polluted by other factors, interfering with our ability to make conclusive statements about whether X causes Y.

Within observational studies, there are two pure types – **cross-sectional observational studies,** which focus on variation across **spatial units** at a single **time unit,** and **time-series observational studies,** which focus on variation within a single spatial unit over multiple time units. There are, in addition, hybrid designs, but for the sake of simplicity we will focus on the pure types.[7] Before we get into the two types of observational studies, we need to provide a brief introduction to observational data.

4.3.1 Datum, Data, Data Set

The word "data" is one of the most grammatically misused words in the English language. Why? Because most people use this word as though it were a singular word when it is, in fact, plural. Any time you read "the data is," you have found a grammatical error. Instead, when describing data, the phrasing should be "the data are." Get used to it: You are now one of the foot soldiers in the crusade to get people to use this word appropriately. It will be a long and uphill battle.

The singular form of the word data is "**datum.**" Together, a collection of datum produces data or a "**data set.**" We define observational data sets by the variables that they contain and the spatial and time units over which they are measured. Political scientists use data measured on a variety of different spatial units. For instance, in survey research, the spatial unit is the individual survey respondent. In comparative U.S. state government studies, the spatial unit is the U.S. state. In international relations, the spatial unit is often the nation. Commonly studied time units are months, quarters, and years. It is also common to refer to the spatial and time units that define data sets as the **data set dimensions.**

Two of the most common types of data sets correspond directly to the two types of observational studies that we just introduced. For instance, Table 4.1 presents a cross-sectional data set in which the time unit is the year 1972 and the spatial unit is nations. These data could be used to test the theory that unemployment percentage (X) → government debt as a percentage of gross national product (Y).

[7] The classic statements of observational studies appeared in 1963 in Donald Campbell and Julian Stanley's seminal work *Experimental and Quasi-experimental Designs for Research*.

Table 4.1. Example of cross-sectional data

Nation	Government debt as a percentage of GNP	Unemployment rate
Finland	6.6	2.6
Denmark	5.7	1.6
United States	27.5	5.6
Spain	13.9	3.2
Sweden	15.9	2.7
Belgium	45.0	2.4
Japan	11.2	1.4
New Zealand	44.6	0.5
Ireland	63.8	5.9
Italy	42.5	4.7
Portugal	6.6	2.1
Norway	28.1	1.7
Netherlands	23.6	2.1
Germany	6.7	0.9
Canada	26.9	6.3
Greece	18.4	2.1
France	8.7	2.8
Switzerland	8.2	0.0
United Kingdom	53.6	3.1
Australia	23.8	2.6

Time-series observational studies contain measures of X and Y across time for a single spatial unit. For instance, Table 4.2 displays a time-series data set in which the spatial unit is the United States and the time unit is months. We could use these data to test the theory that inflation (X) → presidential approval (Y). In a data set, researchers analyze only those data that contain measured values for both the independent variable (X) and the dependent variable (Y) to determine whether the third causal hurdle has been cleared.

4.3.2 Cross-Sectional Observational Studies

As the name implies, a cross-sectional observational study examines a cross section of social reality, focusing on variation between *individual spatial units* – again, like citizens, elected officials, voting districts, or countries – and explaining the variation in the dependent variable across them.

For example, what, if anything, is the connection between the preferences of the voters from a district (X) and a representative's voting behavior (Y)? In a cross-sectional observational study, the strategy that a researcher would pursue in answering this question involves comparing the aggregated

Table 4.2. Example of time-series data

Month	Presidential approval	Inflation
2002.01	83.7	1.14
2002.02	82.0	1.14
2002.03	79.8	1.48
2002.04	76.2	1.64
2002.05	76.3	1.18
2002.06	73.4	1.07
2002.07	71.6	1.46
2002.08	66.5	1.80
2002.09	67.2	1.51
2002.10	65.3	2.03
2002.11	65.5	2.20
2002.12	62.8	2.38

preferences of voters from a variety of districts (X) with the voting records of the representatives (Y). Such an analysis, of course, would have to be observational, instead of experimental, because this particular X is not subject to experimental manipulation. Such an analysis might take place within the confines of a single legislative session, for a variety of practical purposes (such as the absence of turnover in seats, which is an obviously complicating factor).

Bear in mind, of course, that observational studies have to cross the same four casual hurdles as do experiments. And we have noted that, unlike experiments, with their random assignment to treatment groups, observational studies will often get stuck on our fourth hurdle. That might indeed be the case here. Assuming the other three hurdles can be cleared, consider the possibility that there are confounding variables that cause Y and are also correlated with X, which make the X–Y connection spurious. (Can you think of any such factors?) How do cross-sectional observational studies deal with this critical issue? The answer is that, in most cases, this can be accomplished through a series of rather straightforward statistical controls. In particular, beginning in Chapter 9, you will learn the most common social science research tool for "controlling for" other possible causes of Y, namely the multiple regression model. What you will learn there is that multiple regression can allow researchers to see how, if at all, controlling for another variable (like Z) affects the relationship between X and Y.

4.3.3 Time-Series Observational Studies

The other major variant of observational studies is the time-series observational study, which has, at its heart, a comparison over time within a single spatial unit. Unlike in the cross-sectional variety, which examines relationships between variables across individual units typically at a single time point, in the time-series observational study, political scientists typically examine the variation within one spatial unit over time.[8]

For example, how, if at all, do changes in media coverage about the economy (X) affect public concern about the economy (Y)?[9] To be a bit more specific, when the media spend more time talking about the potential problem of inflation, does the public show more concern about inflation, and when the media spend less time on the subject of inflation, does public concern about inflation wane? We can measure these variables in aggregate terms that vary over time. For example, how many stories about inflation make it onto the nightly news in a given month? It is almost certain that that quantity will not be the same each and every month. And how much concern does the public show (through opinion polls, for example) about inflation in a given month? Again, the percentage of people who identify inflation as a pressing problem will almost certainly vary from month to month.

Of course, as with its cross-sectional cousin, the time-series observational study will require us to focus hard on that fourth causal hurdle. Have we controlled for all confounding variables (Z) that are related to the varying volume of news coverage about inflation (X) and public concern about inflation (Y)? (The third exercise at the end of this chapter will ask for your thoughts on this subject.) If we can identify any other possible causes of why the public is sometimes more concerned about inflation, and why they are sometimes less concerned about it, then we will need to control for those factors in our analysis.

4.3.4 The Major Difficulty with Observational Studies

We noted that experimental research designs carry some drawbacks with them. So, too, do observational studies. Here, we focus only on one, but it is a big one. As the preceding examples demonstrate, when we need to control for the other possible causes of Y to cross the fourth causal hurdle, we need to control for *all of them*, not just one.[10] But how do we know whether we have controlled for all of the other possible causes of Y? In many cases, we don't know that for certain. We need to try, of course, to control statistically for all other possible causes that we can, which involves

[8] The spatial units analyzed in time-series observational studies are usually aggregated.

[9] See Iyengar and Kinder (2010).

[10] As we will see in Chapter 9, technically we need to control only for the factors that might affect Y and are also related to X. In practice, though, that is a very difficult distinction to make.

carefully considering the previous research on the subject and gathering as much data on those other causes as is possible. But in many cases, we will simply be unable to do this perfectly.

What all of this means, in our view, is that observational analysis must be a bit more tentative in its pronouncements about causality. Indeed, if we have done the very best we can to control for as many causes of Y, then the most sensible conclusion we can reach, in many cases, is that X causes Y. But in practice, our conclusions are rarely definitive, and subsequent research can modify them. That can be frustrating, we know, for students to come to grips with – and it can be frustrating for researchers, too. But the fact that conclusive answers are difficult to come by should only make us work harder to identify other causes of Y. An important part of being a scientist is that we very rarely can make definitive conclusions about causality; we must remain open to the possibility that some previously unconsidered (Z) variable will surface and render our previously found relationships to be spurious.

4.4 SUMMARY

For almost every phenomenon of interest to political scientists, there is more than one form of research design that they could implement to address questions of causal relationships. Before starting a project, researchers need to decide whether to use experimental or observational methods; and if they opt for the latter, as is common, they have to decide what type of observational study to use. And sometimes researchers choose more than one type of design.

Different research designs help shed light on different questions. Focus, for the moment, on a simple matter like the public's preferences for a more liberal or conservative government policy. Cross-sectional and time-series approaches are both useful in this respect. They simply address different types of substantive questions. Cross-sectional approaches look to see why some individuals prefer more liberal government policies, and why some other individuals prefer more conservative government policies. That is a perfectly worthwhile undertaking for a political scientist: What causes some people to be liberals and others to be conservatives? But consider the time-series approach, which focuses on why the public as an aggregated whole prefers a more liberal or a more conservative government at different points in time. That is simply a different question. Neither approach is inherently better or worse than the other, but they both shed light on different aspects of social reality. Which design researchers should choose depends on what type of question they intend to ask and answer.

CONCEPTS INTRODUCED IN THIS CHAPTER

- aggregate – a quantity that is created by combining the values of many individual cases.
- control group – in an experiment, the subset of cases that is not exposed to the main causal stimulus under investigation.
- correlational studies – synonymous with "observational study."
- cross-sectional observational studies – a research design that focuses on variation across spatial units at a single time unit.
- data set – synonym for "data." A collection of variable values for at least two observations.
- data set dimensions – the spatial and time units that define a data set.
- datum – the singular form of the word data.
- experiments – research designs in which the researcher both controls and randomly assigns values of the independent variable to the participants.
- external validity – the degree to which we can be confident that the results of our analysis apply not only to the participants and circumstances in the study, but also to the population more broadly construed.
- field experiment – an experimental study that occurs in the natural setting where the subjects normally lead their lives.
- internal validity – the degree to which a study produces high levels of confidence about whether the independent variable causes the dependent variable.
- natural experiment – situations in nature that are not properly defined as experiments but the values of the independent variable arise naturally in such a way as to make it seem as if true random assignment by a researcher has occurred.
- observational studies – research designs in which the researcher does not have control over values of the independent variable, which occur naturally; it is necessary that there be some degree of variability in the independent variable across cases, as well as variation in the dependent variable.
- placebo – in an experiment, an innocuous stimulus given to the control group.
- population – the entire set of cases to which our theory applies.
- random assignment – when the participants for an experiment are assigned randomly to one of several possible values of X, the independent variable.
- random sampling – a method for selecting individual cases for a study in which every member of the underlying population has an equal probability of being selected.

- replication – a scientific process in which researchers implement the same procedures repeatedly in identical form to see if the relationships hold in a consistent fashion.
- research designs – the strategies that a researcher employs to make comparisons with the goal of evaluating causal claims.
- sample of convenience – a sample of cases from the underlying population in which the mechanism for selecting cases is not random.
- spatial units – the physical unit that forms the basis for observation.
- survey experiment – a survey research technique in which the interviewing process includes experimental randomization in the survey stimulus.
- time units – the time-based unit that forms the basis for observation.
- time-series observational studies – a research design that focuses on variation within a single spatial unit over multiple time units.
- treatment group – in an experiment, the subset of cases that is exposed to the main causal stimulus under investigation.

EXERCISES

1. Consider the following proposed relationships between an independent and a dependent variable. In each case, would it be realistic for a researcher to perform an experiment to test the theory? If yes, briefly describe what would be randomly assigned in the experiment; if not, briefly explain why not.
   (a) An individual's level of religiosity (X) and his or her preferences for different political candidates (Y)
   (b) Exposure to negative political news (X) and political apathy (Y)
   (c) Military service (X) and attitudes toward foreign policy (Y)
   (d) A speaker's personal characteristics (X) and persuasiveness (Y)
2. Consider the relationship between education level (X) and voting turnout (Y). How would the design of a cross-sectional observational study differ from that of a time-series observational study?
3. In the section on time-series observational studies, we introduced the idea of how varying levels of media coverage of inflation (X) might cause variation in public concern about inflation (Y). Can you think of any relevant Z variables that we will need to control for, statistically, in such an analysis, to be confident that the relationship between X and Y is causal?
4. In the previous chapter (specifically, the section titled "Why Is Studying Causality So Important? Three Examples from Political Science"), we gave examples of research problems. For each of these examples, identify the spatial unit(s) and time unit(s). For each, say whether the study was an experiment, a cross-sectional observational study, or a time-series observational study.

5. Table 4.1 presents data for a test of a theory by use of a cross-sectional observational study. If this same theory were tested by use of a time-series observational study, what would the data table look like?
6. Compare the two designs for testing the preceding theory. Across the two forms of observational studies, what are the Z variables for which you want to control?
7. Table 4.2 presents data for a test of a theory by use of a time-series observational study. If this same theory were tested by use of a cross-sectional observational study, what would the data table look like?
8. Compare the two designs for testing the preceding theory. Across the two forms of observational studies, what are the Z variables for which you want to control?
9. Use your library's resources or Google Scholar (`scholar.google.com`) to look up the following articles and determine whether the research design used in each is an experiment, a cross-sectional observational study, or a time-series observational study. (Note: To access these articles, you might need to perform the search from a location based on your campus.)
   (a) Clarke, Harold D., William Mishler, and Paul Whiteley. 1990. "Recapturing the Falklands: Models of Conservative Popularity, 1979–83." *British Journal of Political Science* 20(1):63–81.
   (b) Gibson, James L., Gregory A. Caldeira, and Vanessa A. Baird. 1998. "On the Legitimacy of National High Courts." *American Political Science Review* 92(2):343–358.
   (c) Druckman, James N. 2001. "The Implications of Framing Effects for Citizen Competence." *Political Behavior* 23(3).

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 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.

[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.

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 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

[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.

[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 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

[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.

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.

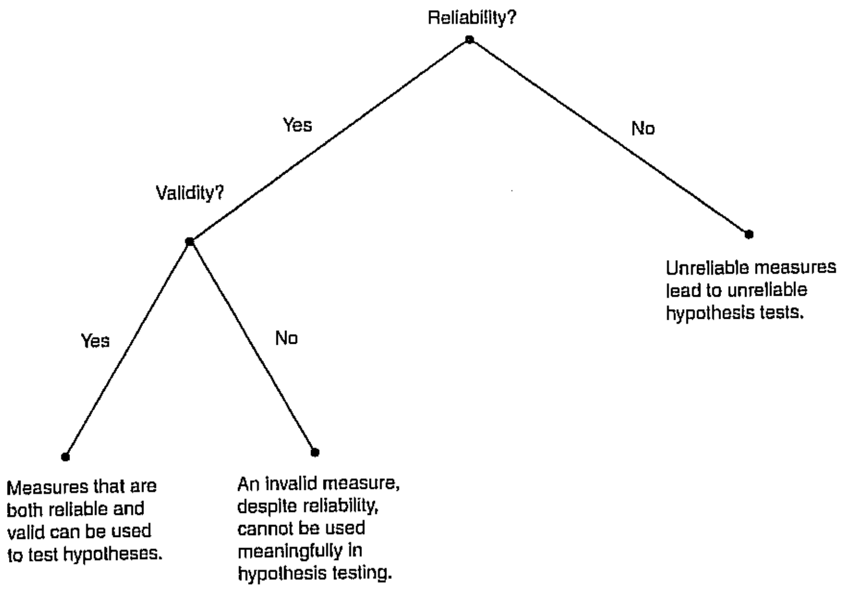

Reliability?
Yes
No
Validity?
Unreliable measures lead to unreliable hypothesis tests.
Yes
No
Measures that are both reliable and valid can be used to test hypotheses.
An invalid measure, despite reliability, cannot be used meaningfully in hypothesis testing.

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.

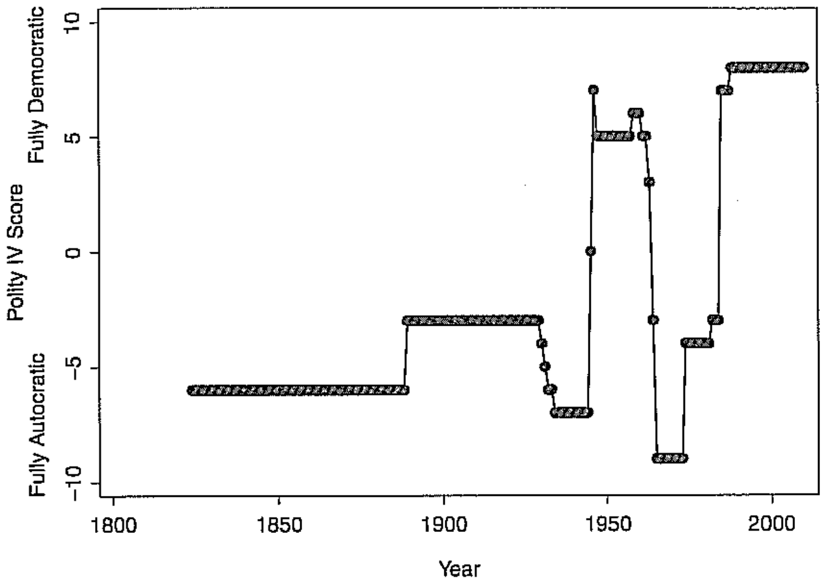


Figure 5.2. Polity IV score for Brazil.

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

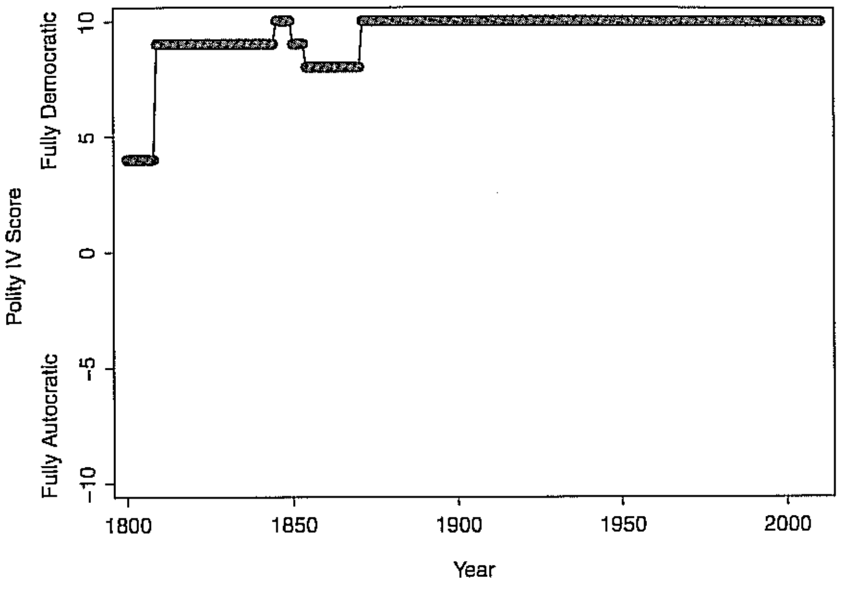


Figure 5.3. Polity IV score for the United States.

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 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.

[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.

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 “Religious Identification” from the NES survey measured during the 2004 national elections in the United States.

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

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.

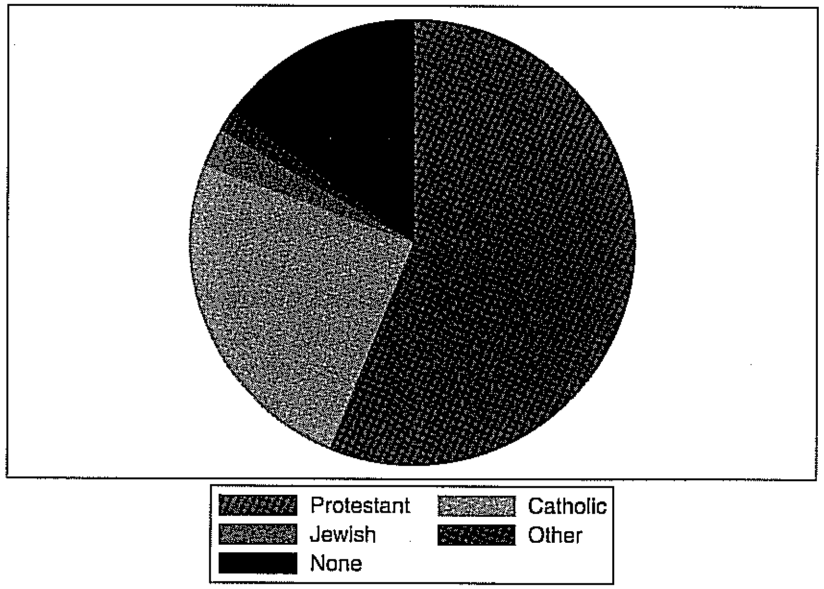


Figure 5.4. Pie graph of religious identification, NES 2004.

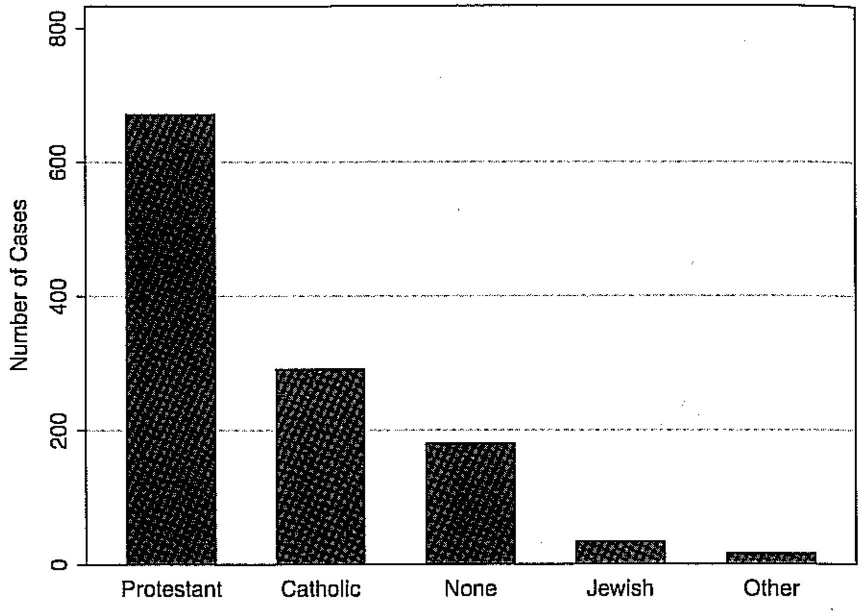
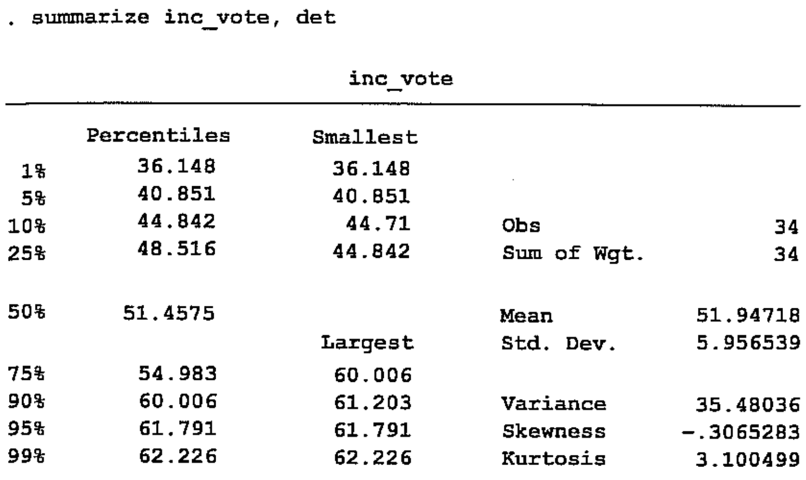


Figure 5.5. Bar graph of religious identification, NES 2004.

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



```
. 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.

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

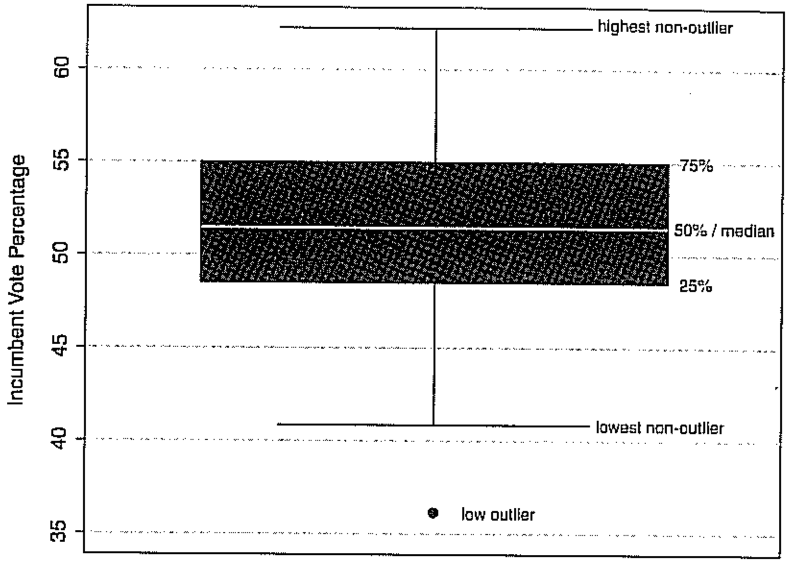


Figure 5.7. Box-whisker plot of incumbent-party presidential vote percentage, 1876–2008.

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

[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.

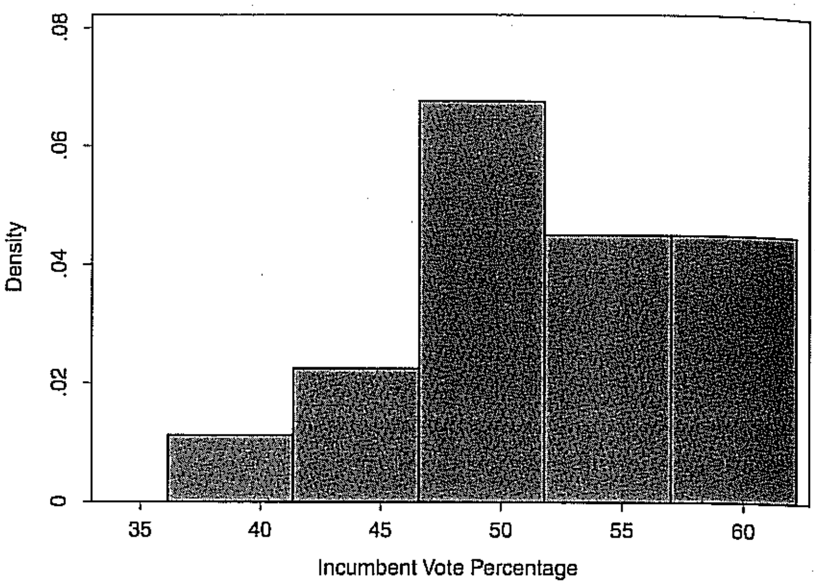


Figure 5.8. Histogram of incumbent-party presidential vote percentage, 1876–2008.

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.

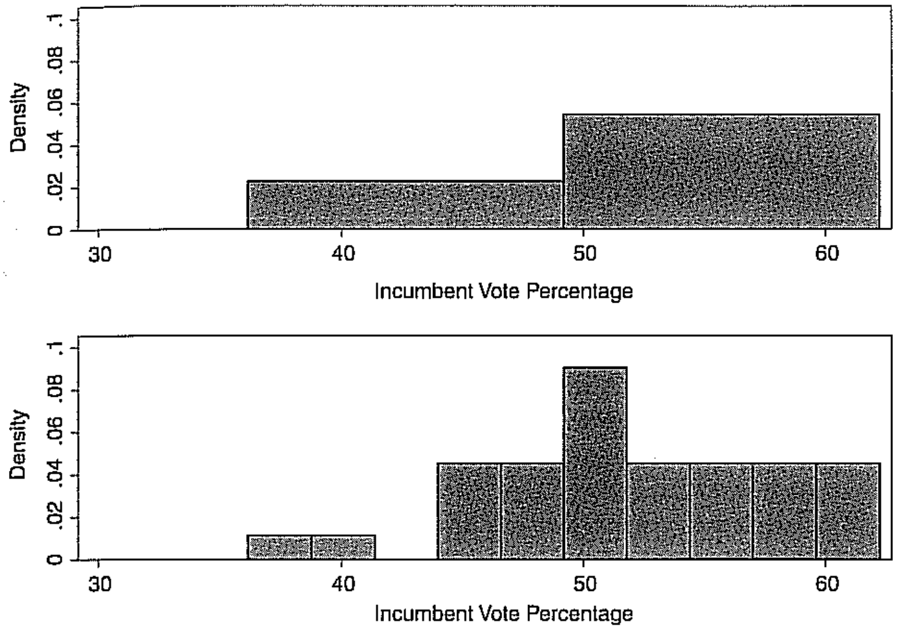


Figure 5.9. Histograms of incumbent-party presidential vote percentage, 1876–2008, depicted with 2 and then 10 blocks.

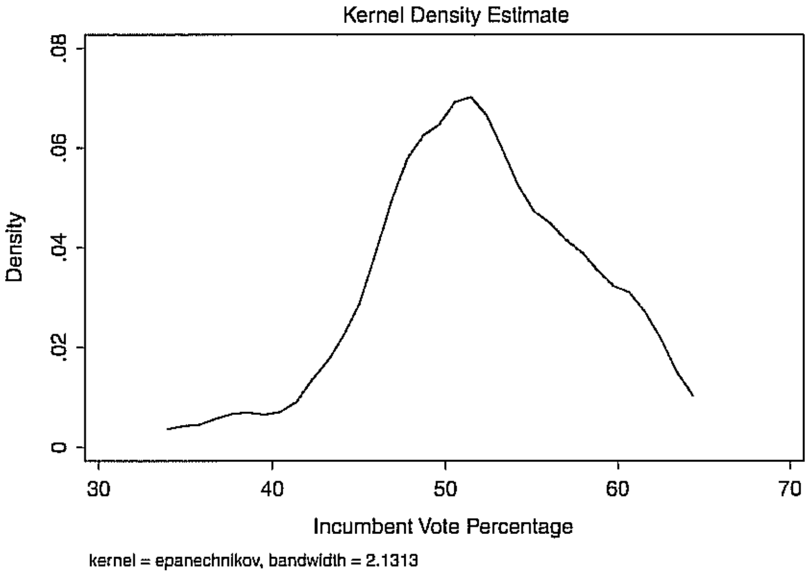


Figure 5.10. Kernel density plot of incumbent-party presidential vote percentage, 1876–2008.

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.

6 Probability and Statistical Inference

OVERVIEW

Researchers aspire to draw conclusions about the entire population of cases that are relevant to a particular research question. However, in most cases, they must rely on data from only a sample of those cases to do so. In this chapter, we lay the foundation for how researchers make inferences about a population of cases while only observing a sample of data. This foundation rests on probability theory, which we introduce here with extensive references to examples. We conclude the chapter with an example familiar to political science students – namely, the "plus-or-minus" error figures in presidential approval polls, showing where such figures come from and how they illustrate the principles of building bridges from samples we know about with certainty to the underlying population of interest.

How dare we speak of the laws of chance? Is not chance the antithesis of all law?
– Bertrand Russell

6.1 POPULATIONS AND SAMPLES

In Chapter 5, we learned how to measure our key concepts of interest, and how to use descriptive statistics to summarize large amounts of information about a single variable. In particular, you discovered how to characterize a distribution by computing measures of central tendency (like the mean or median) and measures of dispersion (like the standard deviation or IQR). For example, you can implement these formulae to characterize the distribution of income in the United States, or, for that matter, the scores of a midterm examination your professor may have just handed back.

But it is time to draw a critical distinction between two types of data sets that social scientists might use. The first type is data about the