CHAPTER 4

CONTENTS

The Challenge of Descriptive Inference

In Chapter 1, we introduced the idea of descriptive inference. We noted that inference is "the process of using the facts we know to learn about facts we do not know," and we divided inferences into two types: descriptive and causal.[1] As the term suggests, in making descriptive inferences, our goal is to describe. This might mean determining what something is, establishing how prevalent or common a phenomenon is, or resolving if it is increasing or decreasing over time. For example, we might want to know: What are the voter turnout rates across different U.S. states? What percent of Afghanis supported the NATO-led military intervention in their country in 2001? Is corruption in African countries increasing or decreasing? Descriptive inference can be contrasted with causal inference, which goes a step further and asks *why* something occurs. Why is voter turnout higher in Midwestern states than in Southern states? Why did some Afghanis support the NATO-led intervention and others did not? Why is corruption increasing in some countries and decreasing in others? As we noted, however, we cannot make causal inferences until we are confident in our descriptive inferences. For example, it would not make much sense to ask why corruption is increasing in Africa, if it is actually stable or even decreasing. As we can see, making descriptive inferences is an important research goal in its own right, and it is also an essential first step to making causal inferences. In this chapter, therefore, we will focus on description.

1 Gary King, Robert O. Keohane, and Sidney Verba, *Designing social inquiry: Scientific inference in qualitative research* (Princeton: Princeton University Press, 1994), p. 46.

In Chapter 1 we introduced three broad challenges to making descriptive inferences: (1) conceptualization, (2) measurement, or operationalization, and (3) case selection, or sampling. In this chapter we pick up these threads and explore these themes in greater detail. The chapter first explores the challenges to defining the concepts that we are interested in measuring and studying. Before moving directly into a discussion on measurement, the chapter explores how the difficulties that we face vary based on the type of data that we are studying. For example, a comparative study of countries produces very different challenges and opportunities than a survey of the U.S. public. Then we explore the challenges of measurement and sampling using different types of data. Having examined the major challenges, the chapter turns towards some of the basic, practical tools available to students and researchers to draw descriptive inferences. These range from basic bar charts and measures of central tendency, used with quantitative data, to narratives and quotes, used with qualitative data.

CONCEPTUALIZATION

The first step in drawing a valid inference is to be clear about just what it is we are making inferences about. Defining our variables of interest is known as **conceptualization**. Perhaps not surprisingly, most of the variables that we are concerned with in political science are difficult to define. Consider for example the difficult task of defining democracy, human rights, globalization, corruption, and even war. We use these terms commonly in political science courses, and yet they mean very different things to different people. Many other variables are not just difficult, but controversial to define. Virtually any hot topic in U.S. politics involves a conflict over definition. Consider for example defining gun rights, religious freedom, or free speech. Individuals on one side of these issues typically favor very broad definitions while their opponents prefer narrow conceptualizations.

Let's explore a concrete example: in the last few decades scholars have had an interest in trying to measure corruption. Understandably, people want to know just how pervasive corruption is in a given country, how that country compares with other countries, and whether the problem is increasing or decreasing over time. These are all great questions, but, before being able to answer them, scholars first have to define what is meant by "corruption." The non-governmental organization Transparency International, one of the pioneers in the study of corruption, uses the simple definition of "the abuse of entrusted authority for private gain."[2] On the one hand, this definition seems perfect because it is straightforward and because it appears to capture what most of us have in our head when we hear the term "corruption." However, if

2 Transparency International, "The Corruption Perceptions Index" (Berlin: Transparency International, 2010).

we dig a little deeper, we can identify several problems with this definition. For example, who defines abuse? Is it defined by the specific laws of a country, by the culture of a country, or are there conceptions of abuse that can be applied universally across legal systems and across different cultural groups? Likewise, who defines authority? Should corruption be limited to public officials or should abuses of "authority" in a business, a non-profit organization, or even in a family be calculated into a country's measure of corruption? What is private gain? Does private gain require a monetary exchange? If an elected official abuses her office for the benefit of her family, is that private gain? What if she commits abuses to benefit her friends or her political party?

In short, corruption *could* be defined as a universal concept or as something very specific to country and culture. It could be defined very narrowly, as specific behaviors that involve public officials and entail a monetary benefit, or it could be defined very broadly, as any abuses of authority for a wide variety of benefits. Our point is that the term "corruption," a term that students of politics use extensively in daily life, is much more complicated than it first appears.

Here is one of the key points in research where we can clearly see the nonlinearity of the process. Ask yourself which definition will be easier to operationalize, or to measure: (1) a broad concept that varies based on the cultural context and involves a wide array of behaviors and actors or (2) a narrow set of specific behaviors involving specific actors that can be observed in any cultural context? From a measurement perspective, the latter is clearly preferable. With this in mind, when trying to arrive at a definition of their concept of study, many scholars have to already be thinking ahead to what they will actually be able to measure in the real world.

This is a major challenge for many students engaging in empirical research for the first time. In much university course work, students are asked to embrace complexity and nuance rather than to simplify and reduce. In fact, if your class had a group activity and tried to define corruption, we are fairly confident that you would arrive at a very broad definition rather than a narrow one. Making our definition of a concept dependent on our measurements can at times be frustrating. Consider for example the idea of "democracy." If democracy is defined literally as "rule by the people," then few countries would actually qualify as democracies. As a result, most political scientists have defined "democracy" as a form of representative democracy involving free and fair elections. Even this narrowing of the concept begs the question: What is meant by free and fair? Narrow definitions of free and fair elections would likely have to tolerate some abuses of civil liberties, press restrictions, abuses of power, nepotism, and clientelism, all of which do not necessarily match with the idea of "democracy" that we have in our heads. Some scholars attempt to recognize this tension by using the term **operational definition**, meaning a definition that can be measured, or operationalized.[3]

3 Operationalization is the process of moving from a theoretical concept to a measurable variable.

Another approach is to move up and down what Giovanni Sartori referred to as a "ladder of generality."[4] Rather than study "democracy," we could move up the ladder of generality and study "regimes," or we could move down the ladder of generality and study a subtype of democracy, such as "parliamentary democracy."[5] In like fashion, we could study a type of corruption, such as "petty corruption," or relatively small bribes paid to public officials to perform or fail to perform their duties. Once we have arrived at an operational definition, we are now ready to think more specifically about measurement. The first step in this process is considering the enormous amount of variation in the type of measurements that we could develop.

DIFFERENT TYPES OF DATA

It stands to reason that conceptualization and measurement challenges will vary considerably based on the type of information that we are interested in. A natural scientist interested in arsenic contamination in water is going to use a very different set of tools and face a very different set of challenges than a political scientist interested in corruption, the effect of negative campaign advertising, or governing common pool resources. Data can be divided into a number of different categories based on the answers to the following questions:

- What is the unit of analysis?
- What is the level of analysis?
- Do the data cover the entire population or are they based on a sample drawn from a larger population?
- Are the data cross-sectional or longitudinal?
- Are the data qualitative or quantitative?

These terms might not have much meaning to you yet, but we will explore each in turn. The **unit of analysis** is simply what is being studied or compared; in political science research, the units of analysis are typically political actors, political acts, or geographic areas. For example, one might study citizens, households, countries, U.S. states, U.S. or foreign cities, legislation introduced in a legislative body, roll call votes of legislators, laws, newspaper articles, court decisions, or words used in speeches of prominent politicians, to name just a few. Different units of analysis present different challenges. For example, even if a researcher compares all the countries in the world, he or she would still have a limited number of observations—just under 200 depending

4 Giovanni Sartori, "Concept misformation in comparative politics," *American Political Science Review* 64 (1970): 1033–1053.

5 David Collier and Steven Levitsky, "Democracy with adjectives," *World Politics* 49 (1997): 430–451.

on how one defines a country. (Yes, just about everything in political science confronts a definitional problem.)[6] Survey data from a survey of U.S. households, on the other hand, often entail far more observations (typically over 1,000 households) but confronts challenges in ensuring that those households studied are representative of the larger U.S. population.

Comparing a study of countries and a study of households illustrates two more distinctions in the types of data political scientists study. The first of these is the level of analysis. The **level of analysis** refers to the scale of the data, or whether or not they have been aggregated. For example a country is made up of millions of households and households are made up of several individuals. At the micro-level, an individual earns an income. At a slightly higher level of analysis, all the individual incomes in a household can be added to yield the household income. Scaling up yet another level, household income can be combined from throughout a country (along with the income from firms and a few other sources) to derive the Gross National Income.

Moving from the micro-level (e.g. individual) to the macro-level (e.g. country) is known as **aggregation**, and moving in the opposition direction is known as **disaggregation**. Often times we are more interested in aggregated data, particularly for making descriptive inferences. For example, if we conduct a survey of 1,000 Americans about whom they plan to vote for in an upcoming presidential election, it doesn't really tell us much to know that respondent number 342 favors the Republican candidate. Instead, we would rather "aggregate" all the individuals' responses to learn that 51 percent of surveyed Americans favor the Republican candidate. As we will see below, however, aggregated data generate their own challenges.

There is another important difference in a study of countries and a study of households. In a study of countries, it is possible (although often difficult) to collect data for all countries. In a study of households, doing so is extremely rare. Once every ten years the U.S. Census Bureau does attempt to conduct a **census**, or a survey of all U.S. households; however, political scientists do not have this luxury. Instead, political scientists interested in public opinion typically study a **sample**, or subset, of the larger **population**, or universe of subjects. How that sample is selected is an essential challenge to making descriptive inference. If inference is using the facts we know to generalize about the facts that we do not know, then it is essential that our sample be representative of the population that we wish to generalize about.

Data can vary in other ways as well. While some data capture a snapshot at one point in time, what we call **cross-sectional data**, other data include

6 Who defines a state? If we use membership in the United Nations as our definition of a state, then there are 193 member nations and two observer states. If we use external recognition as our criteria but not consensual recognition, then non-member states such as Kosovo and Northern Cyprus could be considered states. If statehood is determined irrespective of external recognition, then Somaliland could be considered a state.

changes over time, known as **longitudinal** or **time series data**. For example, a cross-sectional study might measure the level of corruption across all countries for a given year and then compare across countries. A longitudinal study, however, might compare levels of corruption in one country over time. A third approach, called **panel data**, merges these two types of data and compares all countries over time. Obviously, panel data requires collecting a great deal of data, but several ambitious studies have used the approach to answer a variety of research questions.[7] Survey data can also be cross-sectional or longitudinal. Survey firms often include the same question wording in multiple iterations of a survey, allowing us to observe, for example, presidential approval ratings over time. In some special cases, researchers are actually able to survey the same individuals over time. This type of panel study would allow researchers to not just describe the movements of a president's aggregate approval rating over time, but explain why an individual's evaluation of the president improves or worsens over time.

Many of the examples that have been given thus far have been examples of **quantitative** data, or data that can be given a numerical value. However, an enormous amount of data generated in political science are **qualitative**, or non-numerical. For example, given the difficulties in measuring an often illegal act such as corruption, many researchers have favored a more qualitative approach. For example, researchers studying police corruption might conduct in-depth interviews with high- and low-level police officers, journalists, and heads of civil society organizations involved in policing issues. These interviews typically produce reams of interview transcripts that are often not quantified.

In short, there is a great deal of variation in the types of data used by scholars. As summed up in Table 4.1, data vary based on (1) the unit of analysis—ranging from survey respondents to bills introduced in the legislature to countries, (2) the level of analysis—including micro-, meta-, and macro-level data, (3) whether the data covers an entire population or a sample of that population, (4) whether the data entail multiple observations at one point in time, one observation at several points in time, or multiple observations over multiple time periods, and (5) whether or not the data are qualitative or quantitative.

Consider for a moment all the different combinations of these five categories that exist in the political world. For example, a cross-sectional study of the population of countries, using aggregated quantitative data, would be different from a longitudinal study of a sample of countries using aggregated quantitative data. This would, in turn, be distinct from a panel study of a

7 See for example Michael Alvarez, José Antonio Cheibub, Fernando Limongi, *et al.*, "Classifying political regimes," *Studies in Comparative International Development* 31 (1996): 3–36; and José Antonio Cheibub, Jennifer Gandhi, and James Raymond Vreeland, "Democracy and dictatorship revisited," *Public Choice* 143 (2010): 67–101.

TABLE 4.1 Different Types of Data

Type of data	Examples
Unit of analysis	Individuals, households, countries, U.S. states, U.S. or foreign cities, roll call votes of legislators, laws, newspaper articles, etc. . . .
Level of analysis	Micro-, meta-, macro-
Temporality	Cross-sectional, longitudinal, panel
Coverage	Representative sample, non-representative sample, population
Category	Quantitative, qualitative

sample of households using micro-level quantitative data and yet even more different than a cross-sectional study of political elites using micro-level qualitative data. A given combination of these elements is not necessarily preferable to another; however, each combination does produce its own unique set of opportunities and challenges for measurement and research.

OPERATIONALIZATION AND MEASUREMENT ERROR

Returning to our topic of "corruption," let's explore some of the measurements that have been developed, what operational definitions they use, how they vary along the above-mentioned categories, and what measurement challenges they confront. Despite many challenges, corruption researchers have developed several methods to measure slightly different conceptualizations and/or aspects of corruption.

One popular measure used is Transparency International's Global Corruption Barometer, which simply asks citizens if they have had to pay a bribe in their interactions with government officials, such as police officers or personnel from the water or electrical utility company. This certainly seems like a reasonable means to measure corruption, but what "concept" does such an operationalization actually measure? First, such a survey question would only measure "bribery." Second, it is focused only on public officials. And third, it measures what scholars refer to as petty corruption rather than grand corruption. For example, it would not capture an organized crime leader buying off a high-level police official, a construction contractor bribing an administrator for a contract, or a firm bribing a member of parliament for beneficial legislation. The Barometer can therefore be considered a measurement of a subcategory of corruption further down Sartori's ladder of generality, what is sometimes referred to as "administrative corruption," or "petty corruption."

What type of data is a source like the Global Corruption Barometer?

- The unit of analysis is the individual being surveyed.
- As we are dealing with individuals, the level of analysis is micro-; however, if we wanted to know what percentage of respondents in a given country had paid a bribe, then the data could be aggregated to the macro-level.
- If the survey is only conducted one time, then it would be considered cross-sectional; however, because this survey has been conducted almost every year since 2003, it can also be analyzed as longitudinal data.
- The data are based on a sample of a larger population of interest.
- Because respondents are asked to answer "yes" or "no" to questions about bribery, the data can be quantified.

Knowing this basic information about our data allows us to better assess what types of challenges we are likely to face in making descriptive inferences. Let's first consider measurement error. Imagine that our survey asks respondents: "In the last twelve months, have you or anyone living in your household paid a bribe in any form to a government official?" This is a variation of what appears in the Global Corruption Barometer. What potential measurement errors might result from such a survey question?

First, this question is likely to engender **social desirability bias**—that is, the tendency of survey respondents to give a socially desirable response rather than an honest one when asked sensitive questions. In this example, some people might not be comfortable admitting to having paid a bribe, so the survey runs the risk of under-reporting the true amount of bribe payments occurring in a society—introducing **systematic measurement error**, or **bias**. The error is *systematic* because it will consistently underestimate the amount of bribe payments. Scholars often refer to systematic errors as producing **validity** concerns, as the bias might invalidate the measure. Social desirability bias can also produce over-reporting of some behaviors. For example, when asked in surveys if they voted in a presidential election, a large percentage of non-voters will report that they actually voted.[8] This is clearly evidenced by comparing survey data from the National Election Survey with actual voter turnout, the former of which shows considerably higher turnout than the latter.

As you can probably imagine, social desirability bias creates an obstacle to measuring any number of subjects, and, therefore, pollsters have to carefully consider the way questions are worded. In the case of bribery, a researcher might instead ask: "In the last twelve months, has a public official *solicited* a bribe from you?" In this case, a survey respondent answering in the affirmative would be admitting no wrong-doing, reducing the risk of bias.

Beyond social desirability bias, there are other ways that systematic error can find its way into survey data. Suppose you were interested in determining

8 Allyson L. Holbrook and Jon A. Krosnick, "Measuring voter turnout by using the randomized response technique: Evidence calling into question the method's validity," *Public Opinion Quarterly* 74 (2010): 328–343.

how a corruption scandal had impacted a political office holder's approval rating, you would not want to ask, "Following the recent corruption scandal, do you approve or disapprove of the way political officer holder X is handling his job?" By priming the respondents to think of the corruption scandal, this question wording would likely bias the responses. Poor question ordering can create the same problem. For example, consider a survey that primes respondents with a series of questions about governmental corruption, waste, and mismanagement, and then asks respondents if they approve of how the government is performing.

While asking respondents if a public official has *solicited* a bribe might reduce systematic measurement error, the question still risks some **random measurement error**. Random error also causes a measurement to deviate from the concept being studied; however, it does not do so in a predictable way. To illustrate, one could imagine that a given survey respondent had a bribe solicited from him over a year and a half ago, but at the time of the survey he remembered incorrectly and stated that, yes, he had a bribe solicited from him in the last twelve months. One could also imagine a different respondent that had a bribe solicited eight months prior, but he remembered it as having been from a long time ago. Requesting that respondents think back into the past introduces error into the data because recollections are often unreliable. In other words, some people may mistakenly report more bribe payments while others may report less; the measurement would be **unreliable**, but there is no reason to expect the error to be consistent in one direction or the other.

There are several additional ways that a survey question could invite unpredictable error. Questions that are too long, too hard to understand, or too ambiguous might invite multiple interpretations. A common mistake in surveys is a **double-barreled question**, or a question that really ask about two things. For example, a survey of citizens using government services might ask: "In your interactions with local government officials, did you have your problem addressed and were you treated well?" In this case the question is asking about both how a citizen was treated and about the outcome of the interaction. It is of course possible to be treated well but without a positive outcome, or to receive a positive outcome but be treated poorly. Such questions are usually easy to identify because they contain the word "and." While question wording can minimize random error, some random measurement error will always exist. Respondents might misunderstand the question, not give it adequate thought, or be in a bad mood that day—all of which might impact their response and introduce error.

In summary, a survey question asking about bribe payments produces a measure of petty corruption with some risk of understating the true level of corruption because of a social desirability bias. The question also invites random error as it asks respondents to recall a year into the past. Both biases and random error can be reduced through question wording; however, some random error is inevitable. Now let's see how the challenges of conceptualization

and systematic and random measurement error manifest themselves with another example.

A very different measurement is produced by another non-governmental organization Global Integrity. Arguing that corruption cannot be effectively measured, "Global Integrity quantitatively assesses the opposite of corruption, that is, the access that citizens and businesses have to a country's government, their ability to monitor its behavior, and their ability to seek redress and advocate for improved governance."[9] Rather than survey citizens about these issues, Global Integrity, based in Washington D.C., hires researchers in each country of study to conduct research and provide responses to over 300 questions, both about the laws on the books and their enforcement. Such a methodology has several advantages. Rather than simply focus on petty corruption, this methodology allows Global Integrity to address a much broader concept. Furthermore, by relying on experts with specialized knowledge, Global Integrity's measure attempts to achieve both breadth and depth.

There is, however, one major problem with this methodology that should be evident. Because different researchers conduct the scoring for different countries, it is almost impossible to ensure that all the researchers are using the same criteria in their evaluations, a problem known as **inter-coder reliability**. Global Integrity offers several examples of this in their methodology white paper. For example, its score card asks experts to determine if "In practice, civil service asset disclosures are audited," but researchers might respond to this question differently if only senior civil servants are audited or if audits are not conducted regularly.[10] The organization attempts to overcome these problems by providing researchers with a great deal of guidance in filling out the scorecards and through a peer review process, whereby experts review the researchers' findings. In 2011, they added the extra step of convening regional peer reviewers to compare several country scores with a regional perspective. These are good steps that reduce the inter-coder reliability problem, but they cannot remove it entirely.

Perhaps the most commonly referenced cross-national measures of corruption come from Transparency International's Corruption Perception Index (CPI) and a similar index of corruption from the World Bank's governance indicators. The term **index** tells us that the final measurement is produced by combining different pieces of data. In this case, these indices are based on surveys of elites, businesspeople, and analysts conducted by risk assessment firms, development banks, and other groups. These sources ask questions about the respondents' perceptions of corruption in a particular country using questions such as, "How widespread do you think bribe taking and corruption are in this country?"

9 Global Integrity, "Global Integrity Report: 2011 Methodology White Paper," Washington, D.C.: Global Integrity. Accessed January 2013. www.globalintegrity.org/report/methodology/white-paper.

10 Global Integrity, "Global Integrity Report: 2011 Methodology White Paper."

As you can see, the CPI uses a very different conceptualization than in the previous examples. It uses a broad conceptualization of corruption and measures "perceptions" rather than actual corruption. In this case, the data are aggregated to the country level and countries are quantitatively ranked on a 1–10 scale. One can identify several measurement concerns with such indices. For example, the data used to calculate these indices come from a variety of different sources, meaning that the questions used or the type of respondents asked might be somewhat different from country to country. In addition, as the index attempts to measure a broader conceptualization of corruption than petty bribery, different respondents might understand corruption differently. Furthermore, because the data are aggregated from individual respondents, it might mask disagreements among respondents regarding their perception of corruption in a given country. To be sure, the end product is attractive: a scale of corruption along which each country in the world can be placed. However, the easy availability of such scales risks blinding the reader to the limitations of the measurement. Consumers of such indices often fail to recognize that there is considerable uncertainty in the data.

The existence of different measurements, using divergent conceptualizations of corruption, and confronting different types of error, has major implications for research findings. This is evident when one compares indices like the CPI with surveys of self-reported bribe payment like Transparency International's Global Corruption Barometer. Figure 4.1 does exactly this. Specifically, the figure, called a scatterplot, places forty-five nations (represented by open circles) on the graph according to its value on the CPI (the *x*-axis) and the Global Corruption Barometer (the *y*-axis). As Figure 4.1 clearly shows, corruption perceptions by elites are very different than self-reported bribe payments by ordinary citizens. While countries that score well on the CPI also generally have fewer self-reported bribe payments, countries that score poorly on the CPI have both high and low levels of bribe payments.

Aggregating data, as done in the Global Corruption Barometer or the CPI, is attractive as it allows us to compare countries; however, as alluded to above, doing so risks masking differences among the individual respondents. In his famous 2006 Technology, Entertainment, Design (TED) Talk, Hans Rosling lays out the problem with aggregation. He notes that, in 2003, Sub-Saharan Africa had an average GDP per capita of only $1,750, clearly the poorest region in the world. Nonetheless, Rosling argues that it is misleading to think about Sub-Saharan Africa as a whole. He notes that, at that time, Sierra Leone only had a GDP per capita of $518 while Mauritius's GDP per capita was an impressive $10,700. Still, country-level data are also an aggregation, so Rosling moves one more step down the ladder of aggregation and divides the data by income quintiles. He notes that after a terrible famine in Niger, the lowest-income quintile (the income for the poorest one-fifth of the country) only had a GDP per capita of $102. During the same time, the highest income

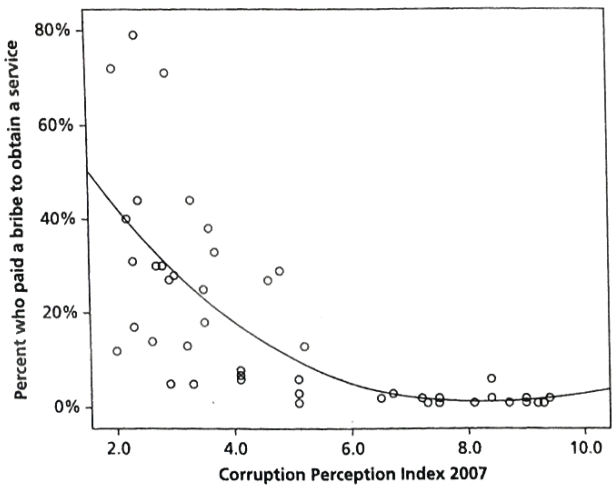

FIGURE 4.1 Scatterplot of the Relationship between Transparency International's CPI and Its Global Corruption Barometer

Source: Author's calculations based on data provided by Transparency International for 2007.
Note: $n = 45$ as the Global Corruption Barometer is only conducted in some countries.

quintile in South Africa had a GDP per capita of $30,400. In response to this dramatic difference he states:

> And yet we tend to discuss on what solutions there should be in Africa. Everything in this world exists in Africa! You can't discuss universal access to HIV [medication] for that quintile up here [South Africa] with the same strategy as down here [Niger]. Improvement of the world must be highly contextualized. And it is not relevant to have it on a regional level. We must be much more detailed.[11]

As you can see, clarity about the level of analysis is particularly important for making descriptive inferences. To offer another example, consider employment rates in the United States. As of December 2012, the unemployment rate in the United States was 7.8 percent. While this is an important piece of information, it is potentially very misleading because it is an aggregation. Among individuals with a bachelor's degree or higher, unemployment was only 3.9 percent compared with 11.6 percent among those with less than a

11 Hans Rosling, "Debunking myths about the third world," *TED Talks*. 2010. Accessed February 8, 2013. www.youtube.com/watch?v = RUwS1uAdUcI.

high school diploma.[12] In short, aggregation is often desirable, but it masks important variation in the data. Having now considered the challenges of conceptualization, measurement error, and aggregation, one last major challenge to descriptive inference remains to be explored: sampling error.

OPERATIONALIZATION AND SAMPLING ERROR

In many cases, particularly when we are dealing with large populations (like all American voters, for example), it is not feasible to collect data on everybody we are interested in. In these situations, we take a subset of the total population, which we refer to as a sample. As alluded to above, research that does not include the entire population to be studied but rather uses a sample to represent the population faces an additional challenge to inference: sampling error. For example, in calculating the percentage of the adult population that supports the 2010 health care law signed by President Obama, pollsters calculate a sample statistic. A December 2012 survey of 1,000 likely voters by Rasmussen Reports, for example, finds that 46 percent view the law favorably while 49 percent have an unfavorable impression.[13] What social scientists really care about, however, is not the percentage of the *sample* that approves of the health care law, the **sample statistic**, but the percentage of the *population*, which is known as the **population parameter**. In order to draw inferences from a sample about the larger population, the sample must be drawn in such a way that it is *representative* of the population. Every sample runs the risk of **sampling error**. This is to say that any sample might vary either systematically or randomly from the true population.

In drawing representative samples, researchers must first and foremost be very clear about the population that they wish to make inferences about. While it is common for news reports to contend that the approval rating of a president represents the views of all Americans,[14] this is rarely the case. Most opinion polls exclude young people under the age of eighteen, homeless people, prisoners, and Americans living abroad. Furthermore, in predicting

12 "Employment status of the civilian population 25 years and over by educational attainment," Bureau of Labor Statistics. Accessed February 8, 2013. www.bls.gov/news.release/empsit.t04.htm.

13 "73% think health care law likely to cost more than projected," Rasmussen Reports. Accessed October 1, 2012. www.rasmussenreports.com/public_content/politics/current_events/healthcare/december_2012/73_think_health_care_law_likely_to_cost_more_than_projected.

14 For example, "A CNN/Opinion Research Corporation survey released Monday indicates that Obama's approval rating among Americans stands at 54 percent, with 45 percent saying they disapprove of the job he's doing as president." CNN Political Unit. 2011. "CNN Poll: Obama's approval rating edges up thanks to foreign policy." CNN. http://politicalticker.blogs.cnn.com/2011/05/30/cnn-poll-obamas-approval-rating-edges-up-thanks-to-foreign-policy/. May 30.

elections, pollsters are not so much concerned with the general population, but those who are likely to go to the polls and actually vote. This distinction can be an important one, as the voting population tends to be older, more affluent, and better informed about politics than the larger adult population. In addition, surveys measuring petty corruption often limit their population to those individuals who have had contact with government officials.

Once the population is clearly defined, we can then consider how to draw a representative sample from that population. As with measurement error, a sample can diverge either systematically or randomly from a population. **Systematic sampling error** typically results from coverage bias or non-response bias. **Coverage bias** occurs when the **sampling frame**, or the group from which the sample is actually drawn, is somehow different from the population.

Say that we are interested in conducting a survey of students' political attitudes on campus. How should we go about collecting a sample of students to represent the population? Perhaps the easiest means to select survey participants would be to post interviewers in frequently traveled locations on campus and conduct face-to-face interviews. Such a sampling method is known as a **convenience sample**, as it selects participants who are conveniently available. We could imagine several problems with such a method. Those students who are on campus more regularly would be more likely to be selected for the sample. Furthermore, if we surveyed during business hours, then many working students who take evening classes would be excluded altogether. In this case, the sampling frame, or the pool from which our sample is drawn (students on campus during business hours), would not be representative of the population we are interested in (all students).

The best means to avoid coverage bias is through random sampling. Instead of a convenience sample, researchers could obtain a list of names of all the students at the university and then develop a means to randomly select names from that list. This could be done by assigning each student a number and then having a computer develop a list of random numbers. Those students whose number was selected by the computer would enter the sample. Regardless of the exact method, the key element of a **random sample** is that each member of the population has an equal probability of being selected for the sample. (Box 4.1 goes into greater detail in how this is done in practice.) As we will see below, a random sample is not only important for avoiding coverage bias, but it is also essential to quantify how uncertain we are about the inferences we make from our samples.

Even if a researcher uses a random sampling technique, he still faces another potential source of systematic sampling error called non-response bias. **Non-response bias** occurs when we cannot collect data from every observation *selected* into our sample. In public opinion surveys, this happens because many individuals who have been selected choose not to participate. The Pew Research Center, a well-regarded polling organization, reported in 2010 that despite at least seven attempts to contact those selected to be in recent Pew

Box 4.1: Sampling in Practice

Developing a random sample of telephone numbers is a little more complicated than randomly selecting from a known list of students, but it follows a similar logic. Surveyors do not know all the available numbers in the U.S., but they do know all the available area codes. They also have estimates of the population living in each area code. If one area code has twice the number of people of another area, then a computer can be told to generate two random telephone numbers from the first area code for every random number from the second. This form of sampling is known as **probability proportionate to size** sampling.[a]

In the early days of polling, surveys in the United States had to be done door to door because many low-income people did not have a telephone. A telephone survey would have produced a biased sample of the American people. Today the sampling challenge is that more and more households have given up a landline telephone for a cell phone. The National Center for Health Statistics estimated that, as of late 2009, 25 percent of households were cell phone only, and this number is increasing dramatically each year.[b] If those households with and without landlines had the same political views, this sampling concern would not be a problem; however, this is clearly not the case. Christian *et al.* found that only 20 percent of a cell phone-only sample identified as Republican compared with 26 percent of a landline sample. Why the difference? Among other differences, Hispanics and young people are more likely than the rest of the population to live in cell phone-only households and less likely to identify as Republicans. For example, of those adults between the ages of twenty-five and twenty-nine, 49 percent lived in cell phone-only households. As a result, since the 2008 presidential election, most major polling companies have incorporated cell phones into their sampling; however, generating random samples of cell phone numbers is more complicated and costly because of legal protections and because cell phone area codes are not as meaningful as landline area codes.[c] Furthermore, individuals who own both a landline and a cell phone have a higher probability of being selected for a sample than someone who just has one.

Outside the United States, sampling concerns vary. While many developed countries use the same approach to sampling as the U.S., in most low-income countries telephone-based sampling would generate a coverage bias problem. As a result, face-to-face interviews remain the predominant means to collect survey data in the developing world. Because there is no list of every resident, however, a computer-generated random selection process cannot be used. Instead, surveyors most commonly employ a method called **cluster sampling**. Imagine a country divided into clusters, something along the lines of U.S. counties. One could develop a list of all the counties in the country and then randomly select counties to be included in the sampling frame. As with telephone survey area

a. Survey firms use slightly different methodologies to select their samples; however, most major firms offer a description of their methodology on their webpage.
b. Leah Christian, Scott Keeter, Kristen Purcell, *et al.*, "Assessing the cell phone challenge to survey research in 2010," Pew Research Center (2010).
c. Paul Lavrakas and Charles Shuttles, "Cell Phone Sampling Summit II statements on accounting for cell phones in telephone survey research in the US" (2005).

codes, the probability of selecting a given county could be proportional to its population. Once counties are randomly selected, surveyors could randomly select neighborhoods within those counties and randomly select homes within those neighborhoods.

In some ways, sampling in developing countries is more accurate than in the United States. ABC/BBC's polling in Afghanistan, for example, obtained a response rate of 95 percent, essentially eliminating the problem of non-response bias so salient in U.S. polling.[d] In other ways, however, challenges remain. In 2007, for example, the Joint United Nations Programme on HIV/AIDS (UNAIDS) estimated that there were 33.2 million people living with HIV. This represented a major decrease from the agency's previous estimate of 39.5 million.[e] The difference, however, was not due to a drop in actual HIV rates, but to a change in methodology to adjust for sampling bias. The data collection method employed in most countries used prenatal care clinics as the primary means of data collection; however, such clinics were more likely to operate in the urban areas where HIV was more prevalent. The result was sampling bias.

d. Gary Langer, "Afghanistan: Where things stand," ABC News (2009). Accessed February 1, 2013. http://abcnews.go.com/PollingUnit/story?id=6787686&page=1.
e. "AIDS epidemic update. Joint United Nations Programme on HIV," UNAIDS, Geneva (2007).

samples, only 5 to 20 percent typically agreed to participate, a statistic known as a **response rate**.[15] If those who are too busy for a survey, opposed to taking surveys, or are otherwise difficult to reach have different political views than those who end up participating in the survey, then the low response rate will create non-response bias. To illustrate this type of systematic error, one could imagine a survey firm calling randomly selected households and asking them if they want to participate in a poll on terrorism. Those who decide to participate would probably have stronger feelings on terrorism, and therefore different attitudes than those who decline, creating a non-response bias.

It is important to emphasize that issues of coverage and non-response bias are not limited to public opinion polling. For example, one might wish to understand how members of parliaments vote on legislation; but what if not all votes are recorded? Indeed, voice votes do occur with great frequency in many parliaments, and their exclusion from studies of roll call voting may affect the conclusions we reach. This is a form of coverage bias—the researcher is interested in understanding how legislators vote on all pieces of legislation, but the sample is limited to bills on which a recorded vote was held. Because voice votes are more likely to be used when there is less party cohesion, a study

15 Lee Rainie, "Internet, broadband, and cell phone statistics," *Pew Internet & American Life Project* 5 (2010).

of roll call votes would over-estimate the degree of party unity in the legislative body.[16]

Cross-national research is often subject to the equivalent of non-response bias due to missing data. While many indicators are available for all countries, data are often missing from low-income countries with inadequate data collection capacity. Since these omissions have a systematic rather than random cause, they may produce biases in the inferences we draw from cross-national comparisons.

Accounting for Error: Quantifying Uncertainty

The discussion above should make clear that most studies run the risk of error: both measurement and sampling, and both systematic and random. To be able to make inferences, therefore, we have to understand the nature of the error in our data and the amount of uncertainty that it introduces. Fortunately, at least in the case of quantitative data, researchers have developed tools to estimate error and to quantify the amount of uncertainty that it introduces. Not surprisingly, it is hard to "quantify" the error in qualitative data, but qualitative research can still benefit from understanding the logic used to do so with quantitative data.

The most straightforward type of error to account for is random sampling error. Leaving aside systematic sampling error for a moment, random sampling error is a function of two factors: the size of the sample and the amount of variation in the data. One could imagine conducting a study of household income in a large city with a wide range of income levels: low to high and everything in between. In this case, if a surveyor only examines a small random sample of say 300 households, it seems likely that his sample would fail to capture all the variation in the population. If, however, the surveyor increases the random sample to 1,000 households, then there is a higher probability that the sample will reflect the population. Alternatively, one could imagine another city where everyone has close to the same level of income. Now a random sample of 300 might be sufficient to adequately capture the population. By taking into account these two components, sample size and variation, researchers are able to estimate the amount of random error in the data: an estimate known as the **standard error**.[17]

16 Clifford J. Carrubba, Matthew Gabel, Lacey Murrah, *et al.*, "Off the record: Unrecorded legislative votes, selection bias and roll-call vote analysis," *British Journal of Political Science* 36 (2006): 691–704.

17 The standard error for interval-level data (such as income) equals the standard deviation (s) divided by the square root of the sample size.

$$std.error = \frac{s}{\sqrt{n}}$$

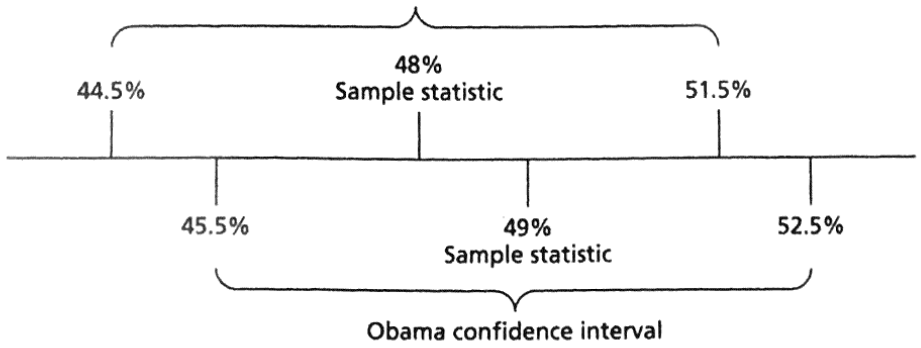

FIGURE 4.2 Confidence Intervals
Source: UPI-CVoter poll, 2012.

Most students first come into contact with the standard error through a term more commonly used in the media: the **margin of error**. For example, in their final poll before the 2012 presidential election, the UPI-CVoter poll of 1,000 likely voters found that 49 percent of likely voters reported an intention to vote for Barack Obama and 48 percent intended to vote for Mitt Romney. The margin of error for the study was reported at 3.5 percent. What that means is that UPI-CVoter was confident that the population parameter, the percentage of the "population" (in this case likely voters in the United States who intended to vote for Barack Obama), was 49 percent plus or minus 3.5 percent, or between 45.5 percent and 52.5 percent (see Figure 4.2). In other words the pollsters found that Obama supporters led Romney supporters in their *sample*, but, when they tried to generalize about the broader *population* of likely voters, they could not be sure of this lead. The one percentage difference observed in the sample statistics could have been due to random chance.

This range of the sample statistic plus and minus the margin of error is known as a **confidence interval** because pollsters are *confident* that the true population parameter, the actual percentage of likely U.S. voters intending to vote for Obama, lies within the range of 45.5 to 52.5 percent. But how confident? The first thing we must stress is that we can never be 100 percent confident; however, pollsters and social scientists want to get as close to 100 percent confidence as is possible. As we said above, the margin of error is related to the standard error. While a margin of error is generally a single measure of uncertainty applied to an entire survey, a standard error is a measure of uncertainty

For sample proportions (such as the percentage of survey respondents who approve of the job

of error is roughly twice the size of the standard error. Thanks to something called the Central Limit Theorem, we know that, 95 percent of the time, a value obtained from a random sample will fall within about two standard errors of the population value. Thus, in our example above based on the UPI-CVoter poll, there is only a 5 percent chance that support for Obama in the population is outside of the 45.5 percent and 52.5 percent range. Another way to think of this is if we conducted 100 surveys of 1,000 likely voters, 95 percent of the time, the sample statistic would be in the 44.5–51.5 percent range.

You might note that this is a pretty large margin of error, which would not be very useful in predicting the outcome of a close electoral race. Nonetheless, as discussed above, pollsters can reduce the uncertainty in the data (reduce the random sampling error) by increasing the sample size. For example, a *Washington Post*–ABC poll conducted right before the election surveyed 2,345 likely voters instead of just 1,000, which reduced the margin of error down to 2 percent.[18] It is important to note, however, that the reduction in the size of the standard error is not consistent. While one can achieve a significant reduction in error when moving from a sample of 500 to one of 1,500, the reduction in error is smaller when moving from 1,500 to 2,500 respondents.[19]

While it is tempting to just focus on the sample statistic and ignore random sampling error in the data, doing so can get us into trouble. Take for example the CPI. Figure 4.3 plots the score for each of the 178 countries rated by Transparency International. The bars on either side of each plot represent 90 percent confidence intervals around each statistic. For example, while Transparency International estimates that Romania scores a 3.7 on its 1–10 scale (where 10 is low corruption), we know that there is random error in this estimate. Taking this random error into account, we are 90 percent confident that the true corruption score falls between 3.3 and 4.2. This is actually a fairly wide range, and this confidence interval overlaps with the intervals for fifty-eight other countries—about one-third of the entire set of countries! Reporting such uncertainty is an extremely important responsibility of researchers,

18 Jon Cohen, Peyton M. Craighill, and Scott Clement, "Wa-Po-ABC tracking poll: Final weekend tally is Obama 50, Romney 47, still a 'margin of error' contest," *Washington Post* (2012). Accessed November 5, 2012. www.washingtonpost.com/blogs/the-fix/wp/2012/11/05/wapo-abc-tracking-poll-final-weekend-tally-is-obama-50-romney-47-still-a-margin-of-error-contest/.

19 Interestingly enough, because today there are so many polls conducted in the lead up to an election, electoral predictions need not rely on any one poll. Websites such as www.huffingtonpost.com/news/pollster/, www.realclearpolitics.com, and http://fivethirtyeight.

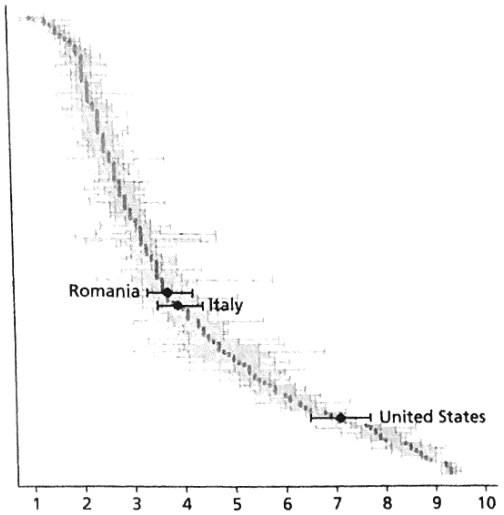

FIGURE 4.3 Plotting Estimates and Uncertainty for the Perception of Corruption Index in 178 Countries

Source: Created by the authors using data from Transparency International.

Note: Bars represent 90 percent confidence intervals.

and paying close attention to such error is an important responsibility of consumers of research.

Researchers and survey analysts use the term **statistically significant difference** to indicate when the differences observed are greater than the error in the data. For example, the confidence interval for the CPI score for the United States runs from 6.5 to 7.7 (See Figure 4.3). Since the lowest end of this range is greater than the high end of the confidence interval for Romania (4.2), we can be more than 90 percent confident that the United States has a higher CPI score than Romania.[20] Alternatively, we cannot be confident that Italy has a higher CPI score than Romania since the confidence interval for Italy overlaps with that for Romania. While it is important to take care to point out when differences we observe are statistically significant, the term needs to be used with caution. For many people the word "significant" implies "substantial," and they are tempted to interpret a "statistically significant difference" as a "large difference." This is incorrect. Statistical significance only suggests that we are

20 In this specific scenario, we would actually be far more confident than 90 percent because the probability is lower that we have underestimated the U.S. score at the extreme of the confidence interval and overestimated the Romanian score at the extreme of the confidence interval.

reasonably confident that there is a difference; however, that difference might actually be very small.[21]

It is also important to recognize that **standard errors and margins of error do not take into account systematic measurement error or systematic sampling error**. We have placed this in bold because it is a common mistake made by both scholars and readers alike. A "statistically significant difference" might be incorrect if there is systematic error in the data.

Taking systematic error into account is more complicated than estimating random error, but it can be done in some cases. To do so, the researcher needs to know certain parameters about the population. Say, for example, a hypothetical polling company measures the attitudes of American adults towards climate change. Having randomly selected its sample, the polling company calculates a sample statistic and the appropriate standard errors. Based on this information, the firm should be confident that a majority of Americans do not view climate change as a priority. However, before going to press with its findings, the company considers the possibility of systematic sampling error. Based on data from the Census Bureau, the firm knows that young people between eighteen and thirty make up about 30 percent of the U.S. adult population. Nonetheless, the pollsters notice that this demographic only makes up 15 percent of their sample. For some reason, perhaps because of their preference for cell phones or a lack of interest in surveys, young people are under-represented in their sample. This would be problematic because their findings also show that young people are more likely to be concerned with global warming.

Fortunately, the pollsters do not have to throw out the data. Instead they can address the systematic error by **weighting** the data. Because they know that young people make up around 30 percent of the population, they can mathematically adjust their sample, or "weight" their data, to look like the population. Researchers can do this by counting each young survey respondent as more than one observation and each person over thirty as less than one observation. Once they have rerun the numbers with the weighted data, the percentage concerned about climate change will increase and the pollster will likely have a more accurate estimation of American adult attitudes towards climate change. Data can be weighted along as many factors as pollsters have population parameters for. The Pew Research Center, for example, notes that it weights its data by household size, combined landline and cell phone users,

21 It should be mentioned that there is some debate about the nature of the error that the standard error actually accounts for. Traditionally, scholars have considered the standard error to be solely a measure of random *sampling* error. Nonetheless, others have noted that some random *measurement* error is also a function of variation and sample size. These researchers contend that the standard error is actually an estimate of all random error: sampling and measurement. Therefore, researchers will often use standard errors and margins of error even when they are studying an entire population. However, from a technical perspective, standard errors and margins of error are calculated in a way so that they are technically only a measure of random sampling error.

age, gender, education, race/ethnicity, and population density. In addition to random sampling, weighting is the reason that U.S. pollsters have been able to accurately capture U.S. public opinion and overcome the problems of non-response bias discussed above (recall the very low response rates cited by the Pew Research Center).

In theory, measurement error can be adjusted in a similar fashion, but again it requires information about the population. We say "in theory," because this practice is far less common in political science. An interesting example can be found in the case of drug consumption. Every year the U.S. Department of Health and Human Services (DHHS) conducts a massive 65,000-person National Survey on Drug Use and Health to measure drug consumption in the United States. Respondents are asked to report on the consumption of various illegal drugs in the past year; however, just as some survey respondents might be hesitant to confess to having paid bribes, some might lie about drug consumption. Recognizing this problem, Kilmer and Pacula compared self-reported drug consumption with more accurate drug tests and estimated that 20 percent of marijuana users fail to report their use of the drug.[22] In a sense, Kilmer and Pucula had developed an estimate of the systematic measurement error in survey data. As a result, when tasked with measuring marijuana consumption in California to predict the impact of a potential legalization initiative, rather than try to administer thousands of drug tests, Kilmer *et al.* adjusted existing survey estimates to account for the known tendency of respondents to under-report their drug use.[23]

In summary, there are means to deal with sampling and measurement error through improved measurements, random sampling, tests of statistical significance, and data weighting, but such tools might not always be readily employable. Moreover, such problems can only be minimized, not eliminated. As social scientists, it is our duty not only to minimize potential threats to inference, but also to report on these threats and to be as specific as possible about how certain or uncertain we are about our inferences. Typically, this involves reporting standard errors and confidence intervals for our estimates, but it may also include discussing how and why our estimates may be biased by sources of systematic error. The reader and the researcher should resist the urge to allow the perceived precision of numeric indicators to generate a false sense of confidence in the resultant numbers.

As mentioned above, there are less clear procedures for estimating and reporting measurement and sampling error in qualitative data; however, the same basic logic can be applied. Imagine that a researcher wanted to conduct a study of corruption in a corruption-prone government agency. It would certainly be important to interview officials from the agency in question, but the researcher

22 Beau Kilmer and Rosalie Liccardo Pacula, *Estimating the size of the global drug market: A demand-side approach, Report 2* (Santa Monica: RAND Corporation, 2009).

23 Beau Kilmer, Jonathan P. Caulkins, Rosalie Liccardo Pacula, *et al.*, *Altered state?* (Santa Monica: RAND Corporation, 2010).

would also have to recognize the potential bias in information provided by such officials. Nonetheless, qualitative researchers also have the luxury of asking more detailed follow-up questions and comparing official answers with those of journalists, members of civil society, and other key informants knowledgeable about the agency. In turn, this interview data can be complemented with other pieces of data—perhaps newspaper articles or citizen complaints filed with an oversight agency. In short, the qualitative researcher is able to use several pieces of information to arrive at an inference, a process often referred to as **triangulation**. By being transparent regarding the method used in arriving at such inferences, by considering what type of random and systematic and measurement and sampling error might be present in the data, and by using wording that recognizes the potential for error, the qualitative researcher can also take error into account.

MAKING DESCRIPTIVE INFERENCES AND PRESENTING DATA

Having discussed several challenges related to conceptualization, measurement, aggregation, and sampling, we are now ready to explore some basic tools used in making descriptive inferences. As the challenges to inference vary with the type of data being analyzed, so too do the tools for descriptive inference. Let's begin with quantitative data and use the concept of "democracy" as an example. Quantitative data can be divided into different levels of measurement, including nominal-, ordinal-, and interval-level data as well as a final level that is far less common in political science research and not addressed here: ratio data (see Figure 4.4).[24]

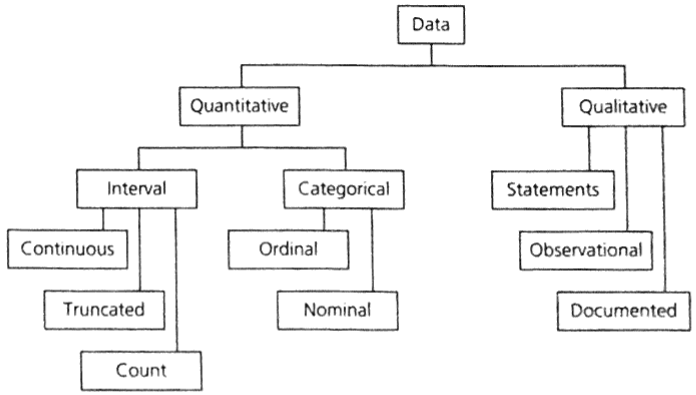

FIGURE 4.4 Different Types of Data

24 Ratio-level data are based on a comparison of two pieces of data to each other: for example the speed of one object compared with another.

TABLE 4.2 Regime Classification

Regime type	Frequency	Percentage	Valid percentage
Parliamentary democracy	56	29.3%	29.6%
Mixed democracy	21	11.0%	11.1%
Presidential democracy	37	19.4%	19.6%
Civilian dictatorship	38	19.9%	20.1%
Military dictatorship	24	12.6%	12.7%
Monarchic dictatorship	13	6.8%	6.9%
Missing data	2	1.0%	
Total	191	191 (100%)	189 (100%)

Source: Democracy Cross-National Data.[25]

Nominal-level data can be divided into different categories, but these categories cannot be placed in an order and the differences between them cannot be described with a precise number. For example, Alvarez *et al.* classify governing regimes into parliamentary democracy (United Kingdom), presidential democracy (Mexico), mixed democracy (France), civilian dictatorship (China), military dictatorship (Equatorial Guinea), and monarchical dictatorship (Saudi Arabia).[26] These are all distinct regime classifications, but it would be difficult to put them in a sequence. One could divide the democracies from the dictatorships, but is a monarchical dictatorship more authoritarian than a military dictatorship? Is a parliamentary system more democratic than a mixed system? Not necessarily. Examples of nominal data in a survey might include religious affiliation, ethnicity, and geographic region of residence.

With nominal-level data we have limited tools of descriptive inference. Useful descriptive statistics include **frequencies**, the number of countries that are classified into each of the six regime types; **percentages**, the percent of the total number of countries, the **valid percent**, the percent of the total number of countries minus any missing data, and the **mode**, or the most common category. This data can be presented visually using either a frequency table (see Table 4.2) or a bar chart (see Figure 4.5).

Ordinal-level data are also divided into set categories, but as its name suggests, these categories can be placed in a sequence. The Freedom House measure of democracy, for example, is often divided into three categories: free,

25 Democracy Cross-National Data, Release 3.0, spring 2009. Accessed August 22, 2013. https://sites.google.com/site/pippanorris3/research/data.

26 Alvarez *et al.*, "Classifying political regimes."

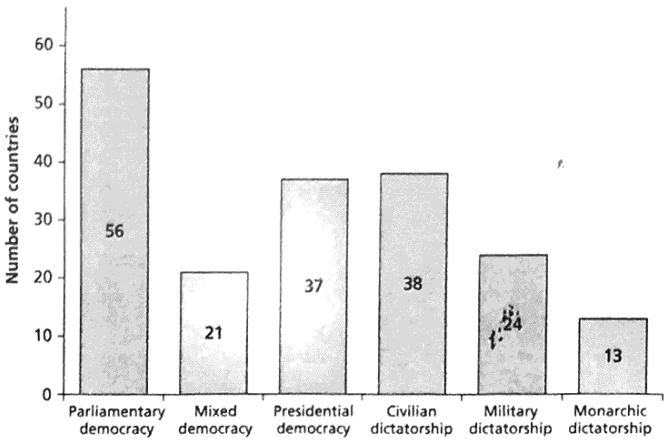

FIGURE 4.5 Nominal-Level Data: Regime Classification across Countries
Source: Democracy Cross-National Data.[27]
Note: n = 189 countries.

partly free, and not free. While these categories are clearly in an order, they do not communicate precise differences. For example, according to the 2011 rankings, Mexico and Kuwait were rated as partly free and Argentina and the United States as free.[28] Such broad categorizations fail to distinguish between the rather small difference in the level of freedom between Argentina and Mexico and the large difference between Kuwait and the United States. This same issue can be seen in survey data. Surveys often ask respondents the extent to which they strongly agree, agree, disagree, or strongly disagree with a given statement. Although these categories can be ordered, the difference between strongly agree and agree might not be the same as the difference between agree and disagree.

Ordinal-level data can be presented with the same descriptive tools, including frequencies, percentages, and the mode. Because the categories can be placed in an order, we are also able to calculate a **median**, or the category that represents the halfway point in the data. In this case the mode is "free" and the median value is "partly free." Figure 4.6 offers a bar chart of the Freedom house rankings for 190 countries.

Bar charts offer a great way to explain data visually, but, if done poorly, graphical representations can be not only confusing, but they also can be

27 Democracy Cross-National Data.

28 Freedom House data along with a white paper explaining the methodology behind their measures is available at www.freedomhouse.org.

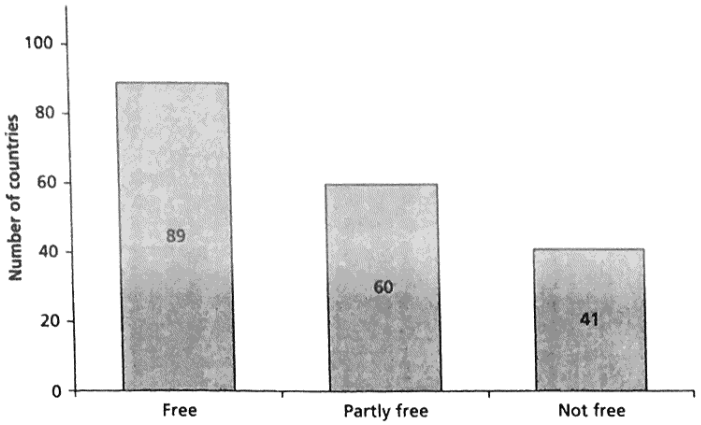

FIGURE 4.6 Ordinal-Level Data: Freedom across Countries

Source: Freedom House 2008.

Note: n = 190 countries.

misleading. Consider the differences between the two bar charts based on the same data presented in Figure 4.7 and Figure 4.8. Figure 4.7 is an example of what not to do.

- The first difference is the scale of the *y*-axis. The smaller scale in Figure 4.7 makes the difference between the two bars look bigger than it actually is, potentially misleading the reader.
- Figure 4.7 also presents the *number* of respondents in each category rather than the *percentage* of respondents. While we do want to know the total sample size, this information can be included in a note. Frequencies are generally more helpful when dealing with smaller sample sizes, and in this case and with most survey data, percentages are clearly more meaningful for the reader.
- Figure 4.8 has clear titles and text and correctly cites where the data came from. Rather than have the reader guess, Figure 4.8 also includes labels specifically stating the value of each bar.
- Figure 4.8 also takes advantage of the fact that many software packages are able to calculate and visually illustrate the confidence intervals around a sample statistic. These error bars clearly show the reader that once we take error into account, the proportions in the population might be a little higher or a little lower than 32 percent and 68 percent.

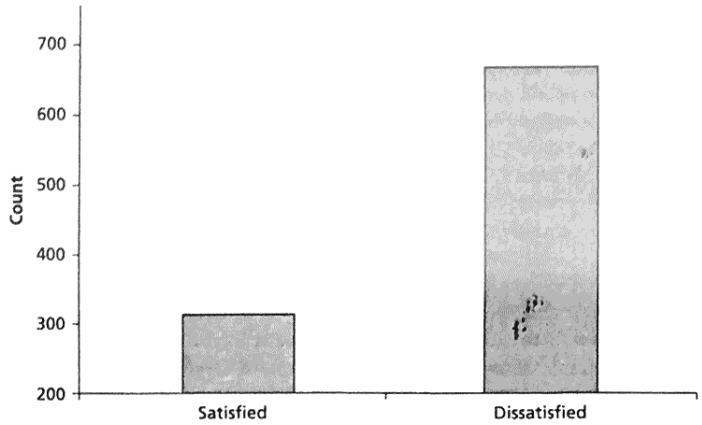

FIGURE 4.7 Example of Poor Data Presentation: Satisfaction with the Way Things Are Going in Egypt, 2010

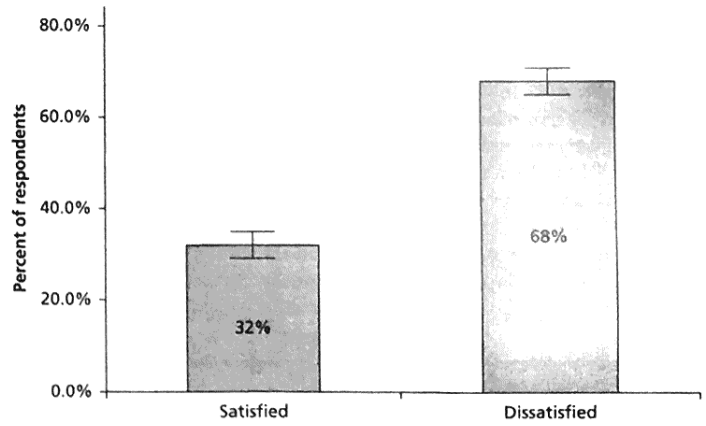

FIGURE 4.8 Example of Good Data Presentation: Percentage "Satisfied with the Way Things Are Going" in Egypt, 2010

Source: Produced by authors using data from the Pew Global Attitudes Project, 2010.

Note: Thin bars represent 95 percent confidence intervals.

The third level of measurement is interval level data, which is both ordered and communicates precise differences. To illustrate, Tatu Vanhanen created an index of democracy based on Robert Dahl's concept of polyarchy, defined by competition and participation.[29] Vanhanen's polyarchy index calculates the percent of seats held by the largest party, as a measure of competition, and electoral turnout, as a measure of participation. He calculates a polyarchy score of 20.78 for Mexico (higher numbers being more democratic) and a score of 26.14 for Argentina, yielding a precise difference of 5.35. Financial indicators are even more intuitive. According to the World Bank, Mexico's GDP per capita in 2011 was estimated at $10,047 and Argentina's was $10,942: a precise difference of $895 per capita.[30] Interval level data can take different forms, for example, some data is **continuous** with no set minimum or maximum. Other data, however, is **truncated**, with a set floor or ceiling. For example, the percent of a country's population that is literate cannot rise over 100 percent. Still other interval level data, such as the age of a survey respondent (if measured in years), is **count data** made up of whole numbers.

Interval level data presents us with a different set of descriptive tools. Frequencies and percentages are generally not particularly helpful with interval level data. For example, there is probably only one country with a GDP per capita of $10,942. Nonetheless scholars have at their disposal a number of other tools, including measures of central tendency and measures related to the distribution or dispersion of the data. We have already mentioned two measures of central tendency, the mode and the median, but for interval level data we can also calculate the mathematical average, or the **mean**. Other measures estimate the distribution of the data. Consider, for example, if we measured income among the inhabitants of two separate islands. On one island everyone earns roughly the same income and on another there are dramatic differences in income. The **standard deviation** allows us a precise measurement of the amount of variation in the data by comparing how far each observation is from the mean.[31] Data from the first island would yield a small standard deviation, with most values close to the mean, and data from the second island would yield a large standard deviation, with many values far away from the mean.

One might also measure the **skewness** in the data or the extent to which the data is lopsided. Income is a case in point. In most communities, there are

29 Vanhanen's democracy index and information about the measurement is available at www.prio.no/CSCW/Datasets/Governance/Vanhanens-index-of-democracy/

30 World Bank data can be easily accessed through its data portal http://data.worldbank.org/.

31 The standard deviation = $\sqrt{\frac{\sum(x_i - \bar{x})^2}{n}}$. The formula simply indicates that to find the standard deviation we must (1) square the difference between the value of x (our variable) for each observation and $\bar{x}$ (the mean for that variable), (2) take the sum of all those squared differences, and then (3) divide by the total number of observations (n), and then (4) take the square root.

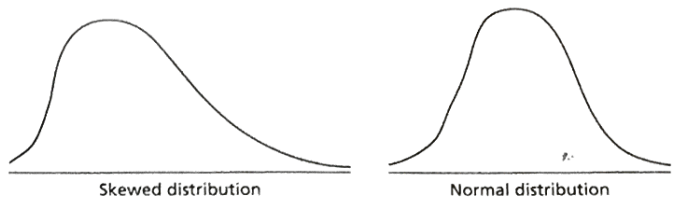

FIGURE 4.9 Skewed and Normal Distributions

large numbers of lower-income and even middle-income individuals but very few high-income individuals (see Figure 4.9). We call this distribution skewed. This distribution of income looks very different than say a distribution of height, where most people are around the mean height and there are a roughly equal number of very short people and very tall people (see Figure 4.9). Such a distribution is known as a **normal distribution**.

Figure 4.10 presents a graphical representation of the CPI across countries. This figure is known as a **histogram**. Because there are probably only one or two countries with a CPI score of say 5.8, a bar chart would not be particularly helpful. Instead a histogram groups values across a range into one bar, or what is sometimes called a bin, and then this bar represents the number

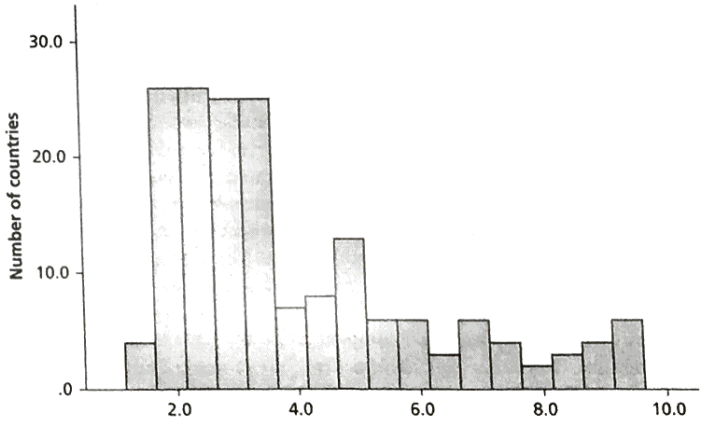

FIGURE 4.10 Example of Histogram: Corruption Perception Index, 2007

Source: Constructed by authors using data from Transparency International.

Note: x-axis represents values on the index of corruption, with low values representing higher levels of corruption; n = 174.

or percentage of countries in this range. In the histogram above, one bar represents a range of approximately .5 on a 1–10 scale. As we can see, the data are very clearly skewed. There are very many high-corruption countries, fewer medium-corruption countries, and only a small minority of low-corruption countries. As a result, while the median is 3.3, the mean is pulled upwards by the low-corruption countries to 4.0. Because the mean is affected by the high-corruption values, we prefer to focus on the median when describing skewed data.

In fact, reporting the mean when we should report the median is an extremely common mistake in describing data. Consider for example the 2012 discussions in the U.S. Congress to avert the so-called "fiscal cliff." One of the issues at hand was the expiration of tax cuts. As it was reported by countless news outlets, the expiration of the tax cuts would cause the average tax payer to pay an additional $3,400 in taxes. While this was mathematically correct, the benefits for the tax cut produced a very skewed distribution. The top 1 percent stood to lose $120,537 in tax cuts while the lowest fifth of households would only pay an additional $412 in taxes. The middle fifth, which contains the median taxpayer and should have been what the media was reporting, stood to lose under $2,000: still a large amount but considerably less than $3,400.[32]

Because interval level data communicates precise differences and includes many different values, there are a host of statistical techniques that we can use with such data. The distinctions between nominal, ordinal, and interval level data offer far more than an interesting classification; the level of measurement determines what statistical techniques we use to describe and present our data. So while we use one set of tools for interval data, we use a very different set of tools for nominal data.

It is a little less clear how to divide qualitative data, but many scholars divide qualitative data into three categories primarily based on the source of the data. These include (1) statements that might be made in a survey, interview, or focus group, (2) observations made by the researcher, or (3) documented information (see Figure 4.4). The presentation of qualitative data can vary considerably. To illustrate, consider for example that we are interested in U.S. citizens' attitudes towards gun control. One qualitative research approach might entail conducting observation and interviews at a protest calling for greater gun control. A researcher could present his data through a narration that describes the protest in depth. Such a narration might entail the use of both quotes from protesters and observations by the researcher to convey the emotions, frustrations, and sentiments of the protesters. Such an account would likely be illuminating, interesting to read, and help the reader understand the viewpoint of the protesters.

32 Rick Newman, "How much the 'fiscal cliff' will cost you," *U.S. News and World Report* (2012). Accessed February 1, 2013. www.usnews.com/news/blogs/rick-newman/2012/11/12/how-much-the-fiscal-cliff-will-cost-you.

Nonetheless, from a point of view of "inference," you should immediately spot the limitation of such a narrative. Individuals attending a gun control protest certainly would not be representative of the broader public or even people sympathetic towards gun control. So how is the qualitative researcher able to respond to this problem?

One option is to be clear about the "population" being studied. Rather than try to make broad generalizations, the goal of the researcher could simply be limited to describing one particular group of protesters. In this case, the researcher would infer from his observation and from interviews about the protesters and the protest itself, but not beyond. This approach has its advantages. The researcher's narrative could provide the reader with a depth of understanding that would be difficult with a survey of gun control attitudes or other quantitatively oriented methods.

The researcher could go a step further and conduct similar research activities at other protests and even at protests defending gun rights. This approach maximizes the strengths of qualitative research. While a survey could give us a sense of general attitudes towards gun control among the population as a whole, observing and interviewing protesters would allow for an analysis of two groups at either end of the political extreme that would otherwise be hard to identify and randomly survey.

Another option would be to use qualitative methods to help better describe quantitative findings. The researcher could use quantitative survey data to broadly describe an issue and then complement this with qualitative data to provide greater depth and understanding. For example, if a survey found that 69 percent of respondents favored a ban on assault weapons, interviews with experts and activists, focus groups, observations at protests, and a review of newspapers or other documents could help explain this 69 percent number in greater detail.

SUMMING UP

There are many cases where we will want to make descriptive inferences. In order to do so scientifically, we have to recognize the numerous challenges to inference that we confront and take steps to minimize these challenges. Based on the discussion above, we can offer the following advice for either conducting your own research or critically analyzing others' research:

- Be very clear about the concept that you want to measure, ensure that the concept can in fact be measured, and confirm that your measurement matches the concept.
- Carefully select the type of data to be collected (e.g. the unit of analysis, qualitative and/or quantitative) and recognize the specific challenges this particular combination of data attributes confronts. For example,

you would want to recognize that aggregating data might mask important differences.

- Select a measurement that minimizes measurement errors. This includes those that produce biases, such as social desirability bias in a survey, and those that produce random measurement error, such as a double-barreled question in a survey.
- If you are not studying the whole population, minimize sampling error through random sampling. Ensure that the sampling frame is constructed to minimize the potential of coverage bias. To the extent possible, reduce non-response bias and missing data.
- Use measurements of random error (e.g. the standard of error and the resultant margin of error) to operationalize uncertainty in the data. Regardless of whether the data are quantitative or qualitative, be transparent about uncertainty in the data.
- When possible, use information available about the population to develop weights and correct for any systematic sampling and (potentially) measurement errors. When not possible, be transparent about any potential systematic error.

Students might not have adequate training to complete these last two steps, but the most important thing is that students and scholars alike are thoughtful about the error that is in their data. They should try to minimize that error whenever possible, and, even when it is not possible, they should be clear about the limitations of the data. Most research papers include a methodology section, where authors should be very specific about any methodological limitations of their study. Valid and reliable descriptive inferences are the first step in ensuring valid and reliable causal inferences. In the following chapters, we will explore three broad approaches to making causal inferences: experiments, large-n observational studies, and small-n observational studies.

KEY TERMS