THE FUNDAMENTALS OF

Political Science Research

Third Edition

Paul M. Kellstedt

Texas A&M University

Guy D. Whitten

Texas A&M University

CAMBRIDGE
UNIVERSITY PRESS

7 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 almost all research situations, 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 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

7.1 POPULATIONS AND SAMPLES

In Chapter 5, we learned how to measure our key concepts of interest, and in Chapter 6 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 **population** – that is, data for every possible relevant case. In your experience, the example of population data that might come to mind first is that of the US Census, an attempt by the US government to gather some critical bits of data about the entire US population once every 10 years.[1] It is a relatively rare occurrence that social scientists will make use of data pertaining to the entire population. But we nevertheless aspire to make inferences about some population of interest, and it is up to the researcher to define explicitly what that population of interest is. Sometimes, as in the case of the US Census, the relevant population – all US residents – is easy to understand. Other times, it is a bit less obvious. Consider a pre-election survey, in which the researcher needs to decide whether the population of interest is all adult citizens, or likely voters, or something else.

The second type of data is drawn from a **sample** – a subset of cases that is drawn from an underlying population. Because of the proliferation of public opinion polls today, many of you might assume that the word "sample" implies a **random sample.**[2] It does not. Researchers *may* draw a sample of data on the basis of randomness – meaning that each member of the population has an equal probability of being selected in the sample. But samples may also be nonrandom, which we refer to as samples of convenience.

The vast majority of analyses undertaken by social scientists are done on sample data, not population data. Why make this distinction? Even though the overwhelming majority of social science data sets are comprised of a sample, not the population, it is critical to note that we are not interested in the properties of the sample *per se*; we are interested in the sample only insofar as it helps us to learn about the underlying population. In effect, we try to build a metaphorical bridge from what we know about the sample to what we believe, probabilistically, to be true about the broader population. That process is called **statistical inference**, because we use what we *know* to be true about one thing (the sample) to *infer* what is likely to be true about another thing (the population). That is what the word "inference" means: It means to draw an uncertain conclusion based on some limited information.

[1] The Bureau of the Census's web site is http://www.census.gov.

[2] When we discussed research design in Chapter 4, we distinguished between the experimental notion of random assignment to treatment groups, on the one hand, and random sampling, on the other. See Chapter 4 if you need a refresher on this difference.

> **YOUR TURN: Two sides of the same coin – "infer" and "imply"**
>
> If a friend of yours revealed *something* to you without saying their true feelings outright, we might say that he or she "implied" something to you. They revealed *some* information, but you had to fill in the gaps to draw the conclusion.
>
> So when your friend "implies" something, in order to fill in those gaps of understanding, you have to "infer" the rest.
>
> Think about how this applies to the concept that we've just defined above – "statistical inference."

There are implications for using sample data to learn about a population. First and foremost is that this process of statistical inference involves, by definition, some degree of uncertainty. That notion, we hope, is relatively straightforward: Any time that we wish to learn something general based on something specific, we are going to encounter some degree of uncertainty. For example, if we want to learn about an entire country's voting-age population, but we don't have the time or resources to interview every member of the voting-age population in that country, we can still learn something about what the population thinks *based upon the observations of a sample of that population*, provided that we know things about how that sample was selected, and provided that we recognize the uncertainty inherent in extrapolating what we know for sure about our sample to what is likely to be true about the population writ large.

In this chapter, we discuss this process of statistical inference, including the tools that social scientists use to learn about the population that they are interested in by using samples of data. Our first step in this process is to discuss the basics of probability theory, which, in turn, forms the basis for all of statistical inference.

7.2 SOME BASICS OF PROBABILITY THEORY

Let's start with an example.

Suppose that you take an empty pillowcase, and that, without anyone else looking, you meticulously count out 550 small blue beads, and 450 small red beads, and place all 1000 of them into the pillowcase. You twist the pillowcase opening a few times to close it up, and then give it a robust shake to mix up the beads. Next, you have a friend reach her hand into the pillowcase – no peeking! – and have her draw out 100 beads, and then count the number of red and blue beads.

Obviously, because she is observing the entire pillowcase and the process unfolding before her eyes, your friend knows that she is taking just a relatively small sample of beads from the population that is in the pillowcase. And because you shook that pillowcase vigorously, and forbade your friend from looking into the pillowcase while selecting the 100 beads, her selection of 100 (more or less) represents a random sample of that population. Importantly, your friend doesn't know, of course, how many red and blue beads are in the pillowcase. She only knows how many red and blue beads she observed in the sample that she plucked out of it.

Next, you ask her to count the number of red and blue beads. Let's imagine that she happened to draw 46 red beads and 54 blue ones. Once she does this, you then ask her the key question: Based on her count, what is her best guess about the *percentage* of red beads and blue beads in the entire pillowcase? The only way for your friend to know for sure how many red and blue beads are in the pillowcase, of course, is to dump out the entire pillowcase and count all 1000 beads. But, on the other hand, you're not exactly asking your friend to make some entirely random, wild guess. She has a decent amount of information, after all, and she can use that information to make a better guess than simply randomly picking a number between 0 and 100 percent.

Sensibly, given the results of her sample, she guesses that 46 percent of the beads in the entire pillowcase are red, and 54 percent are blue.

YOUR TURN: What would you guess in her situation?

Even though you know how many blue and red beads, overall, are in the pillowcase, and you know that your friend's guess based upon her sample is wrong, what should she have guessed, *given the information that she had at the time*?

Before telling her the true answer, you have her dump the 100 beads that she drew back into the pillowcase, re-mix the 1000 beads, and have her repeat the process from the get-go: reach into the pillowcase again, re-draw 100 beads, and count the number of reds and blues drawn again. This time, she draws 43 red beads and 57 blue ones.

You ask your friend if she'd like to revise her guess, and, based on some new information and some quick averaging on her part, she revises her guess to say that she thinks that 44.5 percent of the beads are red, and 55.5 percent of the beads are blue. (She does this by simply averaging the 46 percent of red beads from the first sample and 43 percent of red beads from the second sample.)

The laws of probability are useful in many ways – in calculating gambling odds, for example – but in the above example, they are useful for taking particular information about a characteristic of an observed sample of data and attempting to generalize that information to the underlying (and unobserved) population. The observed samples above, of course, are the two samples of 100 that your friend drew from the pillowcase. The underlying population is represented by the 1000 beads in the bag.

Of course, the example above has some limitations. In particular, in the example, you *knew, with certainty*, the actual population characteristic – there were 450 red and 550 blue beads. In social reality, there is no comparable knowledge of the value of the true characteristic of the underlying population. That's a pretty big difference between our contrived example and the reality that we experience daily.

Now, some definitions.

An **outcome** is the result of a random observation. Two or more outcomes can be said to be **independent outcomes** if the realization of one of the outcomes does not affect the realization of the other outcome. For example, the roll of two dice represents independent outcomes, because the outcome of the first die – did you roll a 1, 2, 3, 4, 5, or 6 – does not affect the outcome of the second die. Rolling a 3 on the first die has no bearing on the outcome of the second die. Hence the outcomes are, by definition, independent, in the sense that one does not depend on the other.

Probability has several key properties. First, all outcomes have some probability ranging from 0 to 1. A probability value of 0 for an outcome means that the outcome is impossible, and a probability value of 1 for an outcome means that the outcome is absolutely certain to happen. As an example of an outcome that cannot possibly happen, consider taking two fair dice, rolling them, and adding up the sides facing up. The probability that the sum will equal 13 is 0, since the highest possible outcome is 12.

Second, the sum of all possible outcomes must be exactly 1. A different way of putting this is that, once you undertake a random observation, you must observe something. If you flip a fair coin, the probability of it landing heads is 1/2, and the probability of it landing tails is 1/2, and the probability of landing either a head or a tail is 1, because $1/2 + 1/2 = 1$.

Third, if (but only if!) two outcomes are independent, then the probability of those events both occurring is equal to the product of them individually. So, if we have our fair coin, and toss it three times, the probability of tossing three tails is $1/2 \times 1/2 \times 1/2 = 1/8$. (Be mindful that each toss is an independent outcome, because seeing a tail on one toss has no bearing on whether the next toss will be a head or a tail.)

Of course, many of the outcomes in which we are interested are not independent. And in these circumstances, more complex rules of probability are required that are beyond the scope of this discussion.

Why is probability relevant for scientific investigations, and in particular, for political science? For several reasons. First, because political scientists typically work with samples (not populations) of data, the rules of probability tell us how we can generalize from what we know with certainty about our sample to what is likely to be true about the broader population. Second, and relatedly, the rules of probability are the key to identifying which relationships are "statistically significant" (a concept that we define in the next chapter). Put differently, we use probability theory to decide whether the patterns of relationships we observe in a sample could have occurred simply by chance.

7.3 LEARNING ABOUT THE POPULATION FROM A SAMPLE: THE CENTRAL LIMIT THEOREM

The reasons that social scientists rely on sample data instead of on population data – in spite of the fact that we care about the results in the population instead of in the sample – are easy to understand. Consider an election campaign, in which the media, the public, and the politicians involved all want a sense of which candidates the public favors and by how much. Is it practicable to take a census in such circumstances? Of course not. The adult population in the United States is well over 200 million people, and it is an understatement to say that we can't interview each and every one of these individuals. We simply don't have the time or the money to do that; and even if we tried, opinion might shift over the time period it would take making the attempt. There is a reason why the US government conducts a **census** only once every ten years.[3]

Of course, anyone familiar with the ubiquitous public opinion polls knows that scholars and news organizations conduct surveys on a sample of Americans routinely and use the results of these surveys to generalize about the people as a whole. When you think about it, it seems a little audacious to think that you can interview perhaps as few as 800 or 1000 people and then use the results of those interviews to generalize to the beliefs and opinions of the entire 200 million. How is that possible?

The answer lies in a fundamental result from statistics called the **central limit theorem**, which Dutch statistician Henk Tijms (2004) calls "the unofficial sovereign of probability theory." Before diving into what the theorem demonstrates, and how it applies to social science research, we need to explore one of the most useful probability distributions in statistics, the **normal distribution**.

7.3.1 The Normal Distribution

To say that a particular distribution is "normal" is *not* to say that it is "typical" or "desirable" or "good." A distribution that is not "normal"

[3] You might not be aware that, even though the federal government conducts only one census per ten years, it conducts sample surveys with great frequency in an attempt to measure population characteristics such as economic activity.

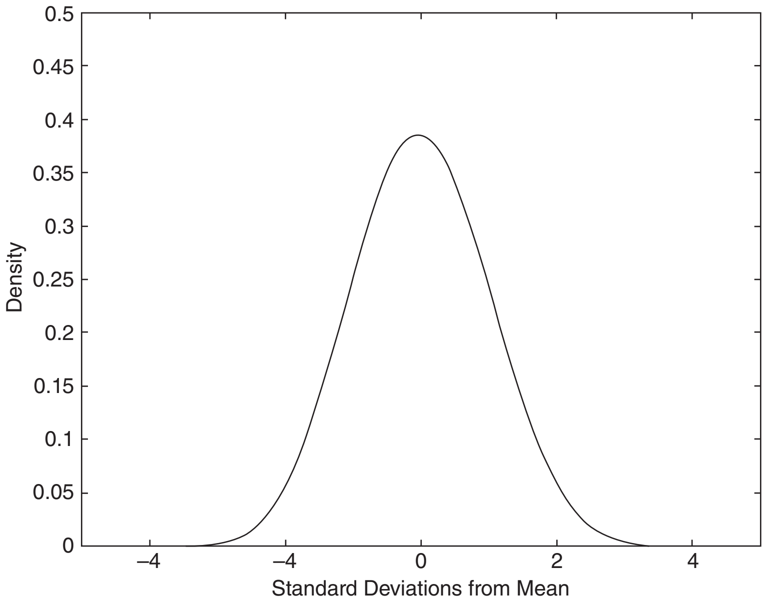

Figure 7.1 The normal probability distribution

is not something odd like the "deviant" or "abnormal" distribution. It is worth emphasizing, as well, that normal distributions are not necessarily common in the real world. Yet, as we will see, they are incredibly useful in the world of statistics.

The normal distribution is often called a "bell curve" in common language. It is shown in Figure 7.1 and has several special properties. First, it is symmetrical about its mean,[4] such that the mode, median, and mean are the same. Second, the normal distribution has a predictable area under the curve within specified distances of the mean. Starting from the mean and going one standard deviation in each direction above and below the mean will capture 68 percent of the area under the curve. Going one additional standard deviation in each direction will capture a shade over 95 percent of the total area under the curve.[5] And going a third standard deviation in each direction will capture more than 99 percent of the total area under the curve. This is commonly referred to as the **68–95–99 rule** and is illustrated in Figure 7.2. You should bear in mind that this is a special feature of the normal distribution and *does not apply to any other-shaped*

[4] Equivalently, but a bit more formally, we can characterize the distribution by its mean and variance (or standard deviation) – which implies that its skewness and excess kurtosis are both equal to zero.

[5] To get exactly 95 percent of the area under the curve, we would actually go 1.96, not 2, standard deviations in each direction from the mean. Nevertheless, the rule of two is a handy rule of thumb for many statistical calculations.

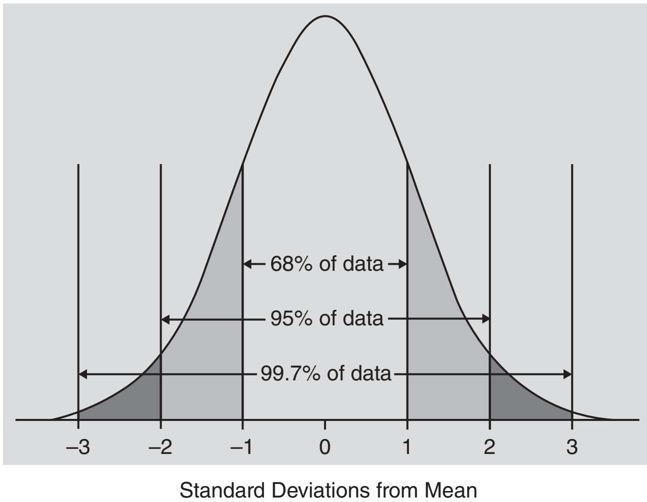

Figure 7.2 The 68–95–99 rule

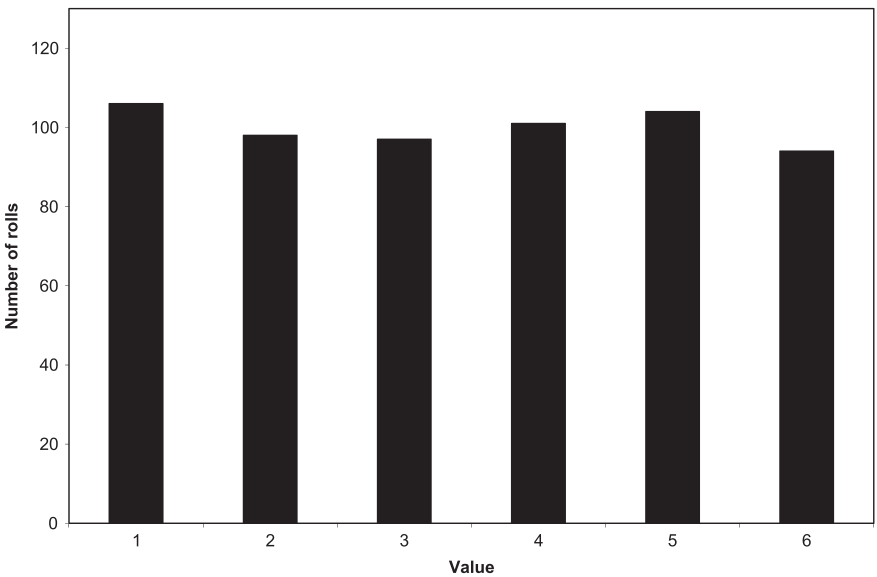

Figure 7.3 Frequency distribution of 600 rolls of a die

distributions. What do the normal distribution and the 68–95–99 rule have to do with the process of learning about population characteristics based on a sample?

A distribution of actual scores in a sample – called a **frequency distribution,** to represent the frequency of each value of a particular variable – on any variable might be shaped normally, or it might not be. Consider the frequency distribution of 600 rolls of a six-sided (and unbiased) die, presented in Figure 7.3. Note something about Figure 7.3

right off the bat: that frequency distribution does not even remotely resemble a normal distribution.[6] If we roll a fair six-sided die 600 times, how many 1s, 2s, etc., should we see? On average, 100 of each, right? That's *pretty close* to what we see in the figure, but only pretty close. Purely because of chance, we rolled a couple too many 1s, for example, and a couple too few 6s.

What can we say about this sample of 600 rolls of the die? And, more to the point, from these 600 rolls of the die, what can we say about the underlying population of all rolls of a fair six-sided die? Before we answer the second question, which will require some inference, let's answer the first, which we can answer with certainty. We can calculate the mean of these rolls of dice in the straightforward way that we learned in Chapter 6: Add up all of the "scores" – that is, the 1s, 2s, and so on – and divide by the total number of rolls, which in this case is 600. That will lead to the following calculation:

$$\begin{aligned}\bar{Y} &= \frac{\sum_{i=1}^{n} Y_i}{n}\\ &= \frac{\sum(1 \times 106) + (2 \times 98) + (3 \times 97) + (4 \times 101) + (5 \times 104) + (6 \times 94)}{600}\\ &= 3.47.\end{aligned}$$

Following the formula for the mean, for our 600 rolls of the die, in the numerator we must add up all of the 1s (106 of them), all of the 2s (98 of them), and so on, and then divide by 600 to produce our result of 3.47.

We can also calculate the standard deviation of this distribution:

$$s_Y = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}} = \sqrt{\frac{1753.40}{599}} = 1.71.$$

Looking at the numerator for the formula for the standard deviation that we learned in Chapter 6, we see that $\sum(Y_i - \bar{Y})^2$ indicates that, for each observation (a 1, 2, 3, 4, 5, or 6) we subtract its value from the mean (3.47), then square that difference, then add up all 600 squared deviations from the mean, which produces a numerator of 1753.40 beneath the square-root sign. Dividing that amount by 599 (that is, $n - 1$), then taking the square root, produces a standard deviation of 1.71.

As we noted, the sample mean is 3.47, but what should we have *expected* the mean to be? If we had exactly 100 rolls of each side of the die, the mean would have been 3.50, so our sample mean is a bit lower than we would have expected. But then again, we can see that we rolled a few "too many" 1s and a few "too few" 6s, so the fact that our mean is a bit below 3.50 makes sense.

[6] In fact, the distribution in the figure very closely resembles a uniform (or flat) distribution.

What would happen, though, if we rolled that same die another 600 times? What would the mean value of those rolls be? We can't say for certain, of course. Perhaps we would come up with another sample mean of 3.47, or perhaps it would be a bit above 3.50, or perhaps the mean would hit 3.50 on the nose. Suppose that we rolled the die 600 times like this not once, and not twice, but an infinite number of times. Let's be clear: We do not mean *an infinite number of rolls*, but instead we mean *rolling the die 600 times for an infinite number of times*. That distinction is critical. We are imagining that we are taking a sample of 600, not once, but an infinite number of times. We can refer to a hypothetical distribution of sample means, such as this, as a **sampling distribution**. It is hypothetical because scientists almost never actually draw more than one sample from an underlying population at one given point in time.

If we followed this hypothetical procedure, we could take those sample means and plot them. Some would be above 3.50, some below, and a few right on it. Here is the key outcome, though: The sampling distribution would be normally shaped, even though the underlying frequency distribution is clearly not normally shaped.

> **YOUR TURN: High and low scores in frequency versus sampling distributions**
>
> With a frequency distribution of 600 rolls of a die that is distributed uniformly, it's not unusual to get extremely high or extremely low rolls of 6 or 1, is it? But if we took the mean of a sample of 600 rolls, what would have to happen in order to get a *sample mean* of 6 or 1?

That is the insight of the central limit theorem. If we can envision an infinite number of random samples and plot our sample means for each of these random samples, those sample means would be distributed normally. Furthermore, the mean of the sampling distribution would be equal to the true population mean. Last, the central limit theorem shows that the standard deviation of the sampling distribution is:

$$\sigma_{\bar{Y}} = \frac{s_Y}{\sqrt{n}},$$

where n is the sample size. The standard deviation of the sampling distribution of sample means, which is known as the **standard error of the mean** (or simply "standard error"), is simply equal to the sample standard deviation of the observed sample divided by the square root of the sample size. In the preceding die-rolling example, the standard error of the mean is

$$\sigma_{\bar{Y}} = \frac{1.71}{\sqrt{600}} = 0.07.$$

Recall that our goal here is to learn what we can about the underlying population based on what we know with certainty about our sample. We know that the mean of our sample of 600 rolls of the die is 3.47, and its standard deviation is 1.71. From those characteristics, we can imagine that, if we rolled that die 600 times an infinite number of times, the resulting sampling distribution would have a standard deviation of 0.07. Our best approximation of the population mean is 3.47, because that is the result that our sample generated.[7] But we realize that our sample of 600 might be different from the true population mean by a little bit, either too high or too low, for no reason other than randomness. What we can do, then, is use our knowledge that the sampling distribution is shaped normally and invoke the 68–95–99 rule to create a **confidence interval** about the likely location of the population mean.

How do we do that? First, we choose a degree of confidence that we want to have in our estimate. Although we can choose any confidence range up from just above 0 to just below 100, social scientists traditionally rely on the 95 percent confidence level. If we follow this tradition – and, critically, because our sampling distribution is normally shaped – we would merely start at our mean (3.47) and move *two* standard errors of the mean in each direction to produce the interval that we are approximately 95 percent confident that the population mean lies within. Why *two* standard errors? Because just over 95 percent of the area under a normal curve lies within two standard errors of the mean. Again, to be precisely 95 percent confident, we would move 1.96, not 2, standard errors in each direction. But the rule of thumb of two is commonly used in practice. In other words,

$$\bar{Y} \pm (2 \times \sigma_{\bar{Y}}) = 3.47 \pm (2 \times 0.07) = 3.47 \pm 0.14.$$

That means, from our sample, we are 95 percent confident that the population mean for our rolls of the die lies somewhere on the interval between 3.33 and 3.61.

YOUR TURN: From 95 to 99 percent confidence intervals

For a variety of reasons, we might like to have more confidence that our estimate lies on a particular interval. Say that, instead of being 95 percent confident, we would be more comfortable with a 99 percent level of confidence. Using what you know about normal distributions, and the procedure we just practiced, can you construct the 99 percent confidence interval?

Is it possible that we're wrong and that the true population mean lies outside that interval? Absolutely. Moreover, we know exactly *how* likely.

[7] One might imagine that our best guess should be 3.50 because, in theory, a fair die ought to produce such a result.

There is a 2.5 percent chance that the population mean is less than 3.33, and a 2.5 percent chance that the population mean is greater than 3.61, for a total of a 5 percent chance that the population mean is not in the interval from 3.33 to 3.61.

Throughout this example we have been helped along by the fact that we knew the underlying characteristics of the data-generating process (a fair die). In the real world, social scientists almost never have this advantage. In the next section we consider such a case.

7.4 EXAMPLE: PRESIDENTIAL APPROVAL RATINGS

Between September 14 and 18, 2017, NBC News and the *Wall Street Journal* sponsored a survey in which 900 randomly selected adult US citizens were interviewed about their political beliefs. Among the questions they were asked was the following item intended to tap into a respondent's evaluation of the president's job performance:

> In general, do you approve or disapprove of the job Donald Trump is doing as president?

This question wording is the industry standard, used for over a half-century by almost all polling organizations.[8] In September of 2017, 43 percent of the sample approved of Trump's job performance, 52 percent disapproved, and 5 percent were unsure.[9]

These news organizations, of course, are not inherently interested in the opinions of those 900 Americans who happened to be in the sample, except insofar as they tell us something about the adult population as a whole. But we can use these 900 responses to do precisely that, using the logic of the central limit theorem and the tools previously described.

To reiterate, we know the properties of our randomly drawn sample of 900 people with absolute certainty. If we consider the 387 approving responses to be 1s and the remaining 513 responses to be 0s, then we calculate our sample mean, $\bar{Y}$, as follows:[10]

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{\sum(387 \times 1) + (513 \times 0)}{900} = 0.43.$$

[8] The only changes, of course, are for the name of the current president.

[9] The source for the survey was http://www.pollingreport.com/djt_job.htm, accessed October 15, 2017.

[10] There are a variety of different ways in which to handle mathematically the 5 percent of "uncertain" responses. In this case, because we are interested in calculating the "approval" rating for this example, it is reasonable to lump the disapproving and unsure answers together. When we make decisions like this in our statistical work, it is very important to communicate exactly what we have done so that the scientific audience can make a reasoned evaluation of our work.

We calculate the sample standard deviation, s_Y, in the following way:

$$s_Y = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}} = \sqrt{\frac{387(1-0.43)^2 + 513(0-0.43)^2}{900-1}}$$
$$= \sqrt{\frac{212.27}{899}} = 0.49.$$

But what can we say about the population as a whole? Obviously, unlike the sample mean, the population mean cannot be known with certainty. But if we imagine that, instead of one sample of 900 respondents, we had an infinite number of samples of 900, then the central limit theorem tells us that those sample means would be distributed normally. Our best guess of the population mean, of course, is 0.43, because it is our sample mean. The standard error of the mean is

$$\sigma_{\bar{Y}} = \frac{0.49}{\sqrt{900}} = 0.016,$$

which is our measure of uncertainty about the population mean. If we use the rule of thumb and calculate the 95 percent confidence interval by using two standard errors in either direction from the sample mean, we are left with the following interval:

$$\bar{Y} \pm (2 \times \sigma_{\bar{Y}}) = 0.43 \pm (2 \times 0.016) = 0.43 \pm 0.032,$$

or between 0.398 and 0.462, which translates into being 95 percent confident that the population value of Trump approval during September 14–18, 2017 was between 39.8 and 46.2 percent.

And this is where the "plus-or-minus" figures that we always see in public opinion polls come from.[11] The best guess for the population mean value is the sample mean value, plus or minus two standard errors. So the plus-or-minus figures we are accustomed to seeing are built, typically, on the 95 percent interval.

7.4.1 What Kind of Sample Was That?

If you read the preceding example carefully, you will have noted that the NBC–*Wall Street Journal* poll we described used a *random* sample of 900 individuals. That means that they used some mechanism (like random-digit telephone dialing) to ensure that all members of the population had an equal probability of being selected for the survey. We want to reiterate the importance of using random samples. The central limit theorem applies

[11] In practice, most polling firms have their own additional adjustments that they make to these calculations, but they start with this basic logic.

only to samples that are selected randomly. With a sample of convenience, by contrast, we cannot invoke the central limit theorem to construct a sampling distribution and create a confidence interval.

This lesson is critical: A nonrandomly selected sample of convenience does very little to help us build bridges between the sample and the population about which we want to learn. This has all sorts of implications about "polls" that news organizations conduct on their web sites. Only certain types of people – high in political interest, with a particular ideological bent – look at these web sites and click on those surveys. As a result, what do such "surveys" say about the population as a whole? Because their samples are clearly not random samples of the underlying population, the answer is "nothing."

There is a related lesson involved here. The preceding example represents an entirely straightforward connection between a sample (the 900 people in the survey) and the population (all adults in the United States). Often the link between the sample and the population is less straightforward. Consider, for example, an examination of votes in a country's legislature during a given year. Assuming that it's easy enough to get all of the roll-call voting information for each member of the legislature (which is our sample), we are left with a slightly perplexing question: What is the population of interest? The answer is not obvious, and not all social scientists would agree on the answer. Some might claim that these data don't represent a sample, but a population, because the data set contains the votes of every member of the legislature. Others might claim that the sample is a sample of one year's worth of the legislature since its inception. Others still might say that the sample is one realization of the infinite number of legislatures that could have happened in that particular year. Suffice it to say that there is no clear scientific consensus, in this example, of what would constitute the "sample" and what would constitute the "population." Still, treating these data as a sample represents the more cautious approach, and hence one we recommend.

7.4.2 Obtaining a Random Sample in the Cellphone Era

In today's era of constant technological advances, you might think that drawing a random sample of adult Americans would be easier than ever. Actually, it's getting harder, and here's why. Let's pretend for the moment that the year is 1988. If a survey-research company wanted to call 900 random US households to conduct a survey, they would use random-digit telephone dialing in a simple way. The phone company (there was only one) told the survey-research company how many numbers were in each three-digit area code. (Some, like (212) for New York City, had more than others, like (402) for the entire state of Nebraska.) Then the phone company would tell you which three-digit prefixes were for households (instead of businesses). And then, armed with those figures, you would call the last four digits randomly. It was not particularly complicated to identify a household. And at the time, over 99 percent of US households had landlines, so very few people would be missed by such a procedure. Lastly, in 1988, the now ubiquitous technology of caller-ID had not yet been made available to households. So a pollster would call a home number, and if someone was home, they picked up, because they didn't know who was calling. In short, random-digit dialing and a little cooperation from the phone company made it easy to get a random sample of the public.

How times have changed! While it is legal for polling companies to call cellphones for survey purposes, it is against US law for those numbers to be auto-dialed by a computer. So the numbers have to be dialed by hand, which takes more time, and costs survey companies more money. Fewer and fewer households – especially younger households, and non-white households – have landlines any more. And caller-ID is everywhere, which means that when you see a number from a polling organization – or simply any number not in your "contacts" list on your cellphone – you might think, "uh, no, I'm not picking up." Or maybe your reaction is the opposite – "Cool, a polling firm! Have I got an opinion for you!" Hopefully you see that even this last point makes pollsters wonder just how "representative" of the population as a whole, or the population of likely voters, their sample of people who answer the phone and agree to be interviewed happens to be.

YOUR TURN: Obtaining a random sample in the 2016 election season

News media organizations have a reputational interest in getting an accurate read of public opinion. And some even try to explain to their audience the process of how they try to guarantee that their samples are indeed representative.

First, go watch the following video from Fox News, and note that the date of the segment is November 30, 2015 – before a single vote was cast during the 2016 primaries or general election: http://video.foxnews.com/v/4638558543001/?#sp=show-clips

Now go watch a video segment from the day after the 2016 general election for an analysis of their polls in the aftermath of the election: http://video.foxnews.com/v/5203990293001/?playlist_id=2127351621001#sp=show-clips

7.4.3 A Note on the Effects of Sample Size

As the formula for the confidence interval indicates, the smaller the standard errors, the "tighter" our resulting confidence intervals will be; larger standard errors will produce "wider" confidence intervals. If we are interested in estimating population values, based on our samples, with as much precision as possible, then it is desirable to have tighter instead of wider confidence intervals.

How can we achieve this? From the formula for the standard error of the mean, it is clear through simple algebra that we can get a smaller quotient by having either a smaller numerator or a larger denominator. Because obtaining a smaller numerator – the sample standard deviation – is not something we can do in practice, we can consider whether it is possible to have a larger denominator – a larger sample size.

Larger sample sizes will reduce the size of the standard errors, and smaller sample sizes will increase the size of the standard errors. This, we hope, makes intuitive sense. If we have a large sample, then it should be easier to make inferences about the population of interest; smaller samples should produce less confidence about the population estimate.

In the preceding example, if instead of having our sample of 900, we had a much larger sample – say, 2500 – our standard errors would have been

$$\sigma_{\bar{Y}} = \frac{0.49}{\sqrt{2500}} = 0.010,$$

which is less than two-thirds the size of our actual standard errors of 0.016. You can do the math to see that going two standard errors of 0.010 in either direction produces a narrower interval than going two standard errors of 0.016. But note that the cost of reducing our error by about 1.2 percent in either direction is the addition of another 1600 respondents, and in many cases that reduction in error will not be worth the financial and time costs involved in obtaining all of those extra interviews.

Consider the opposite case. If, instead of interviewing 900 individuals, we interviewed only 400, then our standard errors would have been

$$\sigma_{\bar{Y}} = \frac{0.49}{\sqrt{400}} = 0.024,$$

which, when doubled to get our 95 percent confidence interval, would leave a plus-or-minus 0.048 (or 4.8 percent) in each direction.

We could be downright silly and obtain a random sample of only 64 people if we liked. That would generate some rather wide confidence intervals. The standard error would be

$$\sigma_{\bar{Y}} = \frac{0.49}{\sqrt{64}} = 0.061,$$

which, when doubled to get the 95 percent confidence interval, would leave a rather hefty plus-or-minus 0.122 (or 12.2 percent) in each direction. In this circumstance, we would guess that Trump's approval in the population was 43 percent, but we would be 95 percent confident that it was between 30.8 and 55.2 percent – and that alarmingly wide interval would be just too wide to be particularly informative.

In short, the answer to the question, "How big does my sample need to be?" is another question: "How tight do you want your confidence intervals to be?"

YOUR TURN: A margin of error of plus-or-minus 1 percent

If the pollsters in the above example for President Trump's approval ratings were willing to tolerate only a margin of error of plus-or-minus 1 percent (for a 95 percent confidence interval), how large would their sample size need to be? Assume, for the moment, that the sample standard deviation remains unchanged at $\sigma_{\bar{Y}} = 0.49$.

7.5 A LOOK AHEAD: EXAMINING RELATIONSHIPS BETWEEN VARIABLES

Let's take stock for a moment. In this book, we have emphasized that political science research involves evaluating causal explanations, which entails examining the relationships between two or more variables. Yet, in this chapter, all we have done is talk about the process of statistical inference with a *single* variable. This was a necessary tangent, because we had to teach you the logic of statistical inference – that is, how we use samples to learn something about an underlying population.

In Chapter 8, you will learn three different ways to move into the world of bivariate hypothesis testing. We will examine relationships between two variables, typically in a sample, and then make probabilistic assessments of the likelihood that those relationships exist in the population. The logic is identical to what you have just learned; we merely extend it to cover relationships between two variables. After that, in Chapter 9, you will learn one other way to conduct hypothesis tests involving two variables – the bivariate regression model.

CONCEPTS INTRODUCED IN THIS CHAPTER

- 68–95–99 rule – a useful characteristic of the normal distribution which states that moving ± 1, ± 2, and ± 3 standard deviations from the mean will leave 68, 95, and 99 percent of the distribution's area under the curve
- census – an accounting of a population
- central limit theorem – a fundamental result from statistics indicating that if one were to collect an infinite number of random samples and plot the resulting sample means, those sample means would be distributed normally around the true population mean