The Essentials of Political Analysis

Third Edition

Philip H. Pollock III
University of Central Florida

CQ PRESS
A Division of SAGE
Washington, D.C.

3. In one of the exercises in Chapter 3 you were asked to consider a study showing that back belts, widely used in industry to prevent injury, do not work. For present purposes, assume that the researchers gathered data on a large number of individual workers. Each worker was measured on a two-category independent variable, labeled "Back belt use." The independent variable's values are: Use belt/Do not use belt. Each worker also was measured on a dependent variable, labeled "Back injury reported." The dependent variable's values are: Reported injury/Did not report injury.[16]
   A. Draw an empty cross-tabulation shell, putting the values of the independent variable on the columns and the values of the dependent variable on the rows. Inside the cross-tabulation, write in fabricated percentages showing that back belt use is not related to back injuries. (Just fabricate percentages. You do not need to fabricate raw cell frequencies.)
   B. According to a report by the Associated Press (December 5, 2000), "The [back belt study's] findings were questioned by a spokesman for the International Mass Retail Association, an industry group whose members include 200 retail chains. The researchers did not directly compare workers doing the same jobs, the spokesman said." The spokesman is suggesting that the original research is flawed because it did not control for the different types of jobs that workers perform. Describe the values of a plausible two-category control variable, labeled "Job type," that measures the types of jobs that workers perform.
   C. The spokesman's claim presents a challenging methodological problem. He is saying that the zero-order relationship, which shows no relationship between back belt use and back injuries, masks a true causal relationship between back belt use and back injuries: Workers who use belts are less likely to report back injuries than workers who do not use belts. The spokesman claims that after controlling for job type, this causal relationship will become evident. (i) Is the spokesman saying that, controlling for job type, the back belt–back injury relationship is spurious? Or is he saying that the back belt–job type–back injury relationships are additive? Or is he saying that interaction is occurring in the back belt–job type–back injury relationships? (ii) Explain how it is possible for the zero-order relationship to show no relationship between belt use and back injuries while at least one of the controlled relationships shows that workers who use belts are less likely to be injured than are workers who do not use back belts.
   D. Draw two cross-tabulation shells, one for each value of job type. Just as you did in part A, put the independent variable on the columns and dependent variable on the rows. Inside the cross-tabulations, write in fabricated percentages that are consistent with your answers in part C.

6 Foundations of Statistical Inference

LEARNING OBJECTIVES

In this chapter you will learn:

- Why random sampling is of cardinal importance in political research
- Why samples that seem small can yield accurate information about much larger groups
- How to figure out the margin of error for the information in a sample
- How to use the normal curve to make inferences about the information in a sample

By this point in the book, you have become comfortable with the essential techniques of political analysis. You know how to think clearly and critically about concepts. You can measure variables, construct explanations, and set up cross-tabulations and mean comparisons. You can interpret complex relationships. As we saw in Chapter 5, however, real-world relationships can present interpretive challenges. For example, suppose that in one of our analyses of the American National Election Study we find that men give the Republican Party an average thermometer rating of 55, compared with a mean rating of 52 among women. Is this 3-point difference "big enough" to support the conclusion that males have higher regard for the Republican Party than do females? Or should we instead decide that the difference is "too small" to warrant that conclusion? Suppose we are investigating the electoral mobilization of military veterans. One of our cross-tabulation analyses shows that 84 percent of veterans reported voting in the presidential election, compared with 77 percent of nonveterans. Does a 7 percentage-point difference allow us to say that veterans are more likely to vote than are nonveterans, or is the difference too fragile to support this implication?

Inferential statistics was invented to help the investigator make the correct interpretations about empirical relationships. **Inferential statistics** refers to a set of procedures for deciding how closely a relationship we observe in a sample corresponds to the unobserved relationship in the population from which the sample was drawn. Inferential statistics can help us decide whether the 3-point feeling thermometer difference between men and women represents a real gender difference in the population or whether the difference occurred by happenstance when the sample was taken. Inferential statistics will tell us how often a random sample will produce a 7 percentage-point difference between veterans and nonveterans if, in fact, no difference exists in the population. In this chapter we cover the essential foundations of inferential statistics. In Chapter 7 we apply these foundational skills to the analysis of empirical relationships.

POPULATION PARAMETERS AND SAMPLE STATISTICS

Anyone who is interested in politics, society, or the economy wants to understand the attitudes, beliefs, or behavior of very large groups. These large aggregations of units are populations. A **population** may be generically defined as the universe of cases the researcher wants to describe. If I were studying the financial activity of political action committees (PACs) in the most recent congressional election, for example, my population would include all PAC contributions in the most recent election. Students analyzing vote choice in the most recent congressional elections, by contrast, would define their population as all voting-age adults. A characteristic of a population—the dollar amount of the average PAC contribution or the percentage of voting age adults who voted—is called a **population parameter.** Figuring out a population's characteristics, its parameters, is a main goal of the social science investigator. Researchers who enjoy complete access to their populations of interest—they can observe and measure every PAC, eligible voter, every member of Congress, Supreme Court decision, or whatever—are working with a **census.** A census allows the researcher to obtain measurements for all members of a population. Thus, the researcher does not need to infer or estimate any population parameters when describing the cases.[1]

More often, however, researchers are unable to examine a population directly and must rely, instead, on a sample. A **sample** is a number of cases or observations drawn from a population. Samples, like death and taxes, are fixtures of life in social research. Because population characteristics are frequently hidden from direct view, we turn to samples, which yield observable sample statistics. A **sample statistic** is an estimate of a population parameter, based on a sample drawn from the population. Public opinion polls, for example, never survey every person in the population of interest (for example, all voting-age adults). The pollster takes a sample, elicits an opinion, and then infers or estimates a population characteristic from this sample statistic. Sometimes such samples—samples of 1,000 to 1,500 are typical—seem too small to faithfully represent their population parameters. Just how accurately does a sample statistic estimate a population parameter? The answer to this question lies at the heart of inferential statistics.

In the sections that follow we will discuss three factors that determine how closely a sample statistic reflects a population parameter. The first two factors have to do with the sample itself: the procedure that we use to choose the sample and the sample's size (the number of cases in the sample). The third factor has to do with the population parameter we want to estimate: the amount of variation in the population characteristic. First we turn to a discussion of the nature and central importance of random sampling. We then consider how a sample statistic, computed from a random sample, is affected by the size of the sample and the amount of variation in the population. Finally, we show how the normal distribution comes into play in helping researchers determine the margin of error of a sample estimate and how this information is used for making inferences.

RANDOM SAMPLING

The procedure we use in picking the sample is of cardinal importance. For a sample statistic to yield an accurate estimate of a population parameter, the researcher must use a **random sample,** that is, a sample that has been randomly drawn from the population. In taking a random sample, the researcher ensures that every member of the population has an equal chance of being chosen for the sample. To appreciate the importance of random sampling, consider a well-known sample that was taken during the 1936 presidential election campaign. Then-president Franklin Roosevelt, a Democrat whose policies were widely viewed as benefiting the lower and working classes, was seeking reelection against Republican candidate Alf Landon, who represented policies more to the liking of higher-income individuals and business interests. In a well-intentioned effort to predict the outcome (and boost circulation), the magazine *Literary Digest* conducted perhaps the largest poll ever undertaken in the history of electoral politics. Using lists of names and addresses obtained from phone records, automobile registrations, and the ranks of its own subscribers, the *Digest* mailed out a staggering 10 million sample ballots, over 2.4 million of which were filled out and returned. Basing its inferences on responses from this enormous sample, the *Digest* predicted a Landon landslide, estimating that 57 percent of the two-party vote would go to Landon and 43 percent to Roosevelt. The election, indeed, produced a landslide—but not for Landon. Roosevelt ended up with more than 60 percent of the vote. (And the *Literary Digest* ended up going out of business.)

What went wrong? In what ways did the magazine's sampling procedure doom its predictions? As you have no doubt surmised, people who owned cars and had telephones (and could afford magazine subscriptions) during the Great Depression may have been representative of Landon supporters, but they decidedly were not a valid reflection of the electorate at large. Certainly, the *Digest* wanted to make a valid inference about the population of likely voters. But it used the wrong **sampling frame,** the wrong method for defining the population it wanted to study. Poor sampling frames lead directly to selection bias, or sampling bias. **Selection bias** occurs when some members of the population are more likely to be included in the sample than are other members of the population. Because people without telephones or cars were systematically excluded from the sample, selection bias was at work. The poll also suffered from **response bias,** which occurs when some cases in the sample are more likely than others to be measured. Because only a portion of the *Digest*'s sample returned their ballots, response bias was at work. People who are sufficiently motivated to fill out and return a sample ballot—or any sort of voluntary-response questionnaire—may hold opinions that are systematically different from the opinions of people who receive a ballot but fail to return it.[2] Samples drawn in this manner are guaranteed to produce sample statistics that are meaningless. Garbage in. Garbage out.

Fortunately, thanks in part to lessons learned from legendary mistakes like the *Literary Digest* poll, social science has figured out how to construct sampling frames that virtually eliminate selection bias and has devised sampling procedures that minimize response bias. A valid sample is based on **random selection.** Random selection occurs when every member of the population has an equal chance of being included in the sample. So, if there are 1,000 members of the population, then the probability that any one member would be chosen is 1 out of 1,000. Thus, the *Literary Digest* should have defined the population they wanted to make inferences about—the entire voting-age population in 1936—and then taken a random sample from this population. By using random selection, every eligible voter, not just those who owned cars or had telephones, would have had an equal chance of being included. But the *Digest*, probably believing that a huge sample size would do the trick, ignored the essential principle of random selection: If a sample is not randomly selected, then the size of the sample simply does not matter.

Let's explore these points, using a plausible example. Suppose that a student organization wants to gauge a variety of student political opinions: how students rate the political parties and the institutions of government, whether they have ever volunteered in a political campaign, their ideological leanings, and so on. As a practical matter, the student researchers cannot survey all 20,000 students enrolled at the university, so they decide to take a sample of 100 students. How might the student pollsters obtain a sample that avoids the infamous pitfalls of the *Literary Digest* poll? The group would first define the sampling frame by assigning a unique sequential number to each student in the population, from 00001 for the first student listed in administration records to 20000 for the last listed. No problem so far. But how do the pollsters guarantee that each student has exactly one chance in 20,000 of being sampled? A systematic approach, such as picking every two-hundredth student, would result in the desired sample size (since 20,000/200 = 100), but it would not produce a truly random sample. Why not? Because two students appearing next to each other in the sampling frame would not have an equal chance of being selected.

To obtain a random sample, the researchers would need a list of five-digit random numbers, created by many computer programs. A random number has a certain chaotic beauty. The first digit is randomly generated from the numbers 0–9. The second is randomly generated from 0–9 as well, and so its value is not connected in any way to the first digit. The third digit is completely independent of the first two, and so on, for each of the five digits. Since there is no rhyme or reason to these numbers, the pollsters can begin anywhere on the list, adding to their sample the student having the same number as the first random number, using the second random number to identify the second student, and continuing until a sample of 100 students is reached. (Any random number higher than 20,000 can be safely skipped, since the list has no systematic pattern.) Variants of this basic procedure are used regularly by commercial polling firms, like Gallup, and academically oriented survey centers, such as the University of Michigan's Institute for Social Research.[3]

The essential methodological goodness of random processes has been previously discussed. In Chapter 1 we saw that random error introduces haphazard static into the measurement process. To be sure, random measurement error is not a welcome sight, but it is a mere annoyance compared with the fundamental distortion introduced by systematic measurement error. In Chapter 4 we found that random assignment is the great neutralizer of selection bias in experimental research design. Random assignment ensures that the test group and the control group will not be systematically different in any way, known or unknown, that could affect the dependent variable. If human choice is allowed to enter the assignment process—the investigators choose one subject over another for the test group or a prospective participant chooses the control group instead of the test—then selection bias is onboard. The rationale for random sampling in observational research is identical to its rationale in experimental design. Because the population is beyond empirical view, we take a random sample, which ensures that each population member has an equal chance of being included. Just as random assignment in experimental research eliminates biased differences between the test group and the control group, so does random sampling eliminate biased differences between the population and the sample.

It is important to point out, however, that in eliminating bias we do not eliminate error. In fact, in drawing a random sample, we are consciously introducing **random sampling error.** Random sampling error is defined as the extent to which a sample statistic differs, *by chance*, from a population parameter. Trading one kind of error for another may seem like a bad bargain, but random sampling error is vastly better because we know how it affects a sample statistic, and we fully understand how to estimate its magnitude. Assuming that we are working with a random sample, the population parameter will be equal to the statistic we obtain from the sample, plus any random error that was introduced by taking the sample:

Population parameter = Sample statistic + Random sampling error.

The student researchers want a sample statistic that provides an unbiased estimate of a true population parameter, a characteristic of all students at the university. They eliminate selection bias by taking a random sample. But they know that random sampling error is affecting their estimate of the population parameter. Assume that the student researchers use a feeling thermometer scale to measure the sample's attitudes toward the Democratic Party. Having collected this information on each member of the sample, they calculate the mean rating of the Democratic Party. Because they are working with a random sample, the student pollsters know that the sample's mean Democratic rating is the same as the population's mean Democratic rating, plus the random error introduced by taking the sample. What makes random sampling error a "better" kind of error is that we have the statistical tools for figuring out how much a sample statistic is affected by random sampling error.

The magnitude of random sampling error depends on two components: (1) the size of the sample and (2) the amount of variation in the population characteristic being measured. Sample size has an inverse relationship with random sampling error: As the sample size goes up, random sampling error goes down. Variation in the population characteristic has a direct relationship with random sampling error: As variation goes up, random sampling error goes up. These two components—the variation component and the sample size component—are not separate and independent. Rather, they work together, in a partnership of sorts, in determining the size of random sampling error. This partnership can be defined by using ideas and terminology that we have already discussed:

Random sampling error = (Variation component) / (Sample size component).

Before exploring the exact properties of this conceptual formula for random sampling error, consider its intuitive appeal. Notice that "Variation component" is the numerator. This reflects its direct relationship with random sampling error. "Sample size component" is the denominator, depicting its inverse relationship with random sampling error. Return to the student organization example and consider an illustration of how these two components work together. Suppose that in the population of 20,000 students there is a great deal of variation in ratings of the Democratic Party. Large numbers of students dislike the Democrats and give them ratings between 0 and 40. Many students like the Democrats and give them ratings between 60 and 100. Still others give ratings in the middle range, between 40 and 60. So the population parameter the student researchers wish to estimate, Democratic Party thermometer ratings, would have a large variation component. Suppose further that the campus group is working with a small-sized random sample. Thus, the variation component is relatively large and the sample size component is relatively small. Dividing the large variation component by the small sample size component would yield a large amount of random sampling error. Under these circumstances, the organization could not be very confident that their sample statistic provides an accurate picture of the true population mean, because their estimate contains so much random sampling error. But notice that if the campus group were to take a larger sample, or if student ratings of the Democratic Party were not so spread out, random sampling error would diminish, and the student pollsters would gain confidence in their sample statistic.

Both components, the variation component and the sample size component, have known properties that give the researcher a good idea of just how much random sampling error is contained in a sample statistic.

Sample Size and Random Sampling Error

As previously noted, the basic effect of sample size on random sampling error is: As the sample size increases, error decreases. Adopting conventional notation—in which sample size is denoted by a lowercase n—we would have to say that a sample of $n = 400$ is preferable to a sample of $n = 100$, since the larger sample would provide a more accurate picture of what we are after. However, the inverse relationship between sample size and sampling error is nonlinear. Even though the larger sample is four times the size of the smaller one, going from $n = 100$ to $n = 400$ delivers only a twofold reduction in random sampling error. In ordinary language, if you wish to cut random error in half, you must quadruple the sample size. In mathematical language, the sample size component of random sampling error is equal to the square root of the sample size, n:

$$\text{Sample size component of random sampling error} = \sqrt{n}.$$

Plugging this into our conceptual formula for random sampling error:

$$\text{Random sampling error} = (\text{Variation component}) \,/\, \sqrt{n}.$$

Because of the nonlinear relationship between sample size and random sampling error, samples that seem rather small nonetheless carry an acceptable amount of random error. Consider three samples: $n = 400$, $n = 1{,}600$, and $n = 2{,}500$. The sample size component of the smallest sample size is the square root of 400, which is equal to 20. So, for a sample of this size, we would calculate random sampling error by dividing the variation component by 20. Random sampling error for the next sample would be the variation component divided by the square root of 1,600, which is equal to 40. So, by going from a sample size of 400 to a sample size of 1,600, we can increase the sample size component of random sampling error from 20 to 40. Notice that by increasing the sample size component from 20 to 40, we double the denominator, $\sqrt{n}$. This has a beneficial effect on random sampling error, effectively cutting it by half. Thus, if resources permit, obtaining a sample of $n = 1{,}600$ would be a smart move. Random sampling error for the largest sample would be equal to the variation component divided by the square root of 2,500, which is equal to 50. Boosting the sample size by 900 cases—from 1,600 to 2,500—occasions a modest increase in the sample size component, from 40 to 50. Sophisticated sampling is an expensive undertaking, and survey designers must balance the cost of drawing larger samples against the payoff in precision. For this reason, most of the surveys you see and read about have sample sizes in the 1,500 to 2,000 range, an acceptable comfort range for estimating a population parameter.

Sample size is an important factor in the accuracy of a sample statistic. You now have a better idea of how random sampling error is affected by n. Suppose the campus organization successfully collects its sample ($n = 100$) and computes a sample statistic, mean Democratic rating. Let's say that the sample rates the Democrats at 59, on average. The group wants to know how much random sampling error is contained in this estimate. As we have just seen, part of the sampling error will depend on the sample size. In this case, the sample size component is equal to $\sqrt{100} = 10$. Now what? What does a sample size error component of 10 have to do with the accuracy of the sample mean of 59, the campus group's estimate of the true rating of the Democratic Party in the student population? The answer depends on the second component of random sampling error, the amount of variation in the population characteristic being measured. As we have seen, this connection is direct: As variation in the population characteristic goes up, random sampling error goes up.

To better appreciate how variation in the population parameter affects random sampling error, consider Figure 6-1, which depicts two possible ways that student ratings of the Democrats might be distributed within the student population. First, suppose that Democratic Party ratings are widely dispersed across the student population, as in Panel A of Figure 6-1. There are appreciable numbers of students in every range of the rating scale, from lower to higher, with only a slight amount of clustering around the center of the distribution. Since variation in the population characteristic is high, the variation component of random sampling error is high. A random sample taken from the population would produce a sample mean that may or may not be close to the population mean—it all depends on which cases were randomly selected. Because each student has an equal chance of being chosen for the sample, one sample might pick up a few more students who reside in the upper range of the distribution. Another sample from the same population may randomly choose a few more students from the lower range. In fact, one might draw a very large number of random samples, each one producing a different sample estimate of the population mean. Now, visualize a population like the one depicted in Panel B of Figure 6-1. Notice that the ratings are clustered around a well-defined center, with fewer cases at the extremes of the scale. Since variation in the population characteristic is low, the variation component of random sampling error is low. A random sample taken from the population would produce a sample mean that is close to the population mean. What is more, repeated sampling from the same population would produce sample mean after sample mean that are close to the population mean—and close to each other.

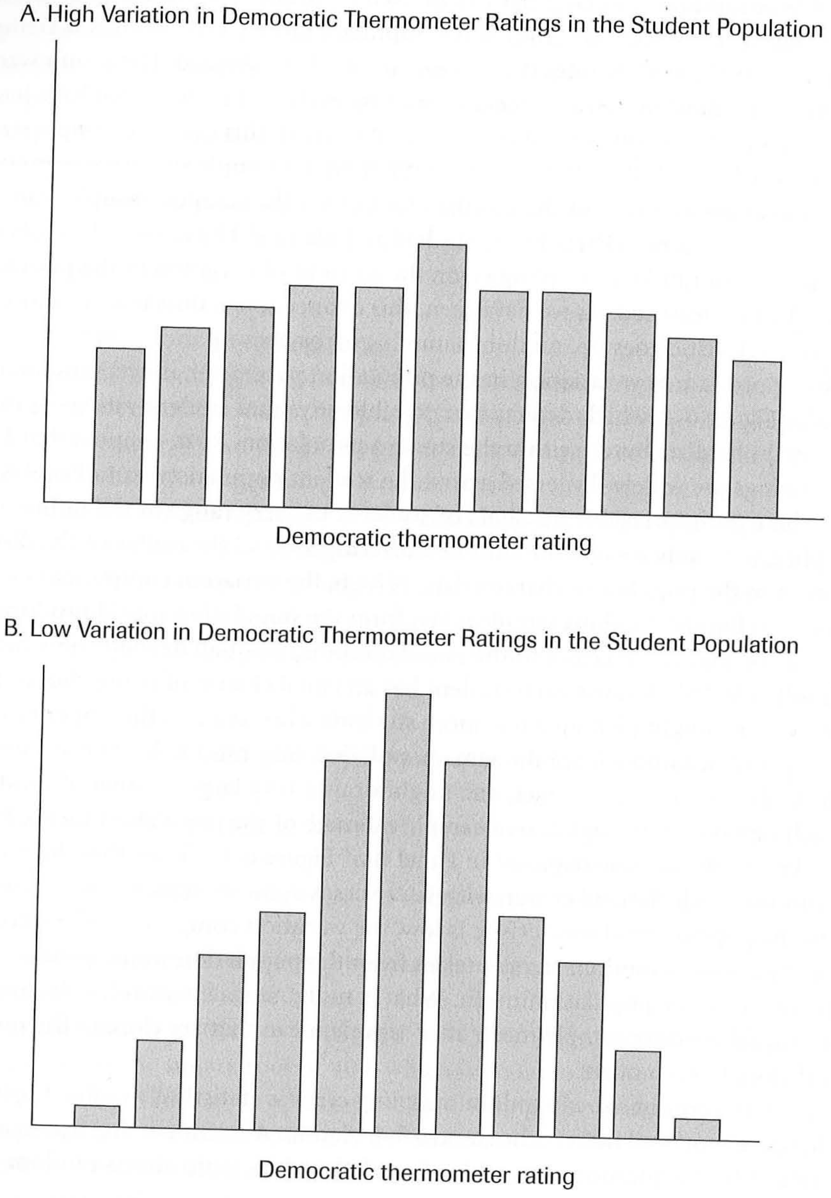
The variation component of random sampling error is statistically defined by a measure you may have encountered before: the standard deviation. After discussing the standard deviation, we return to the question of how this foundational measure affects random sampling error.

Variation Revisited: The Standard Deviation

The amount of variation in a variable is determined by the dispersion of cases across the values of the variable. If the cases tend to fall into one value of a variable, or into a handful of similar values, then the variable has low dispersion. If the cases are more spread out across the variable's values, then the variable has high dispersion. As discussed in Chapter 2, describing the degree of dispersion in nominal and ordinal variables sometimes requires a judgment call.

For interval-level variables, a more precise measure of variation is used. The **standard deviation** summarizes the extent to which the cases in an interval-level distribution fall on or close to the mean of the distribution. Although more precise, the standard deviation is based

Figure 6-1 High Variation and Low Variation in a Population Parameter



A. High Variation in Democratic Thermometer Ratings in the Student Population

B. Low Variation in Democratic Thermometer Ratings in the Student Population

on the same intuition as the less precise judgment calls applied to nominal and ordinal variables. If, on the whole, the individual cases in the distribution do not deviate very much from the distribution's mean, then the standard deviation is a small number. If, by contrast, the individual cases tend to deviate a great deal from the mean—that is, large differences exist between the values of individual cases and the mean of the distribution—then the standard deviation is a large number.

To demonstrate the central importance of the standard deviation in determining random sampling error, we will present two hypothetical possibilities for the distribution of Democratic thermometer ratings in the student population. The population means are the same

Table 6-1 Central Tendency and Variation in Democratic Thermometer Ratings: Hypothetical Scenario A

Student	Democratic rating	Deviation from the mean	Squared deviation from the mean
1	20	–38	1,444
2	22	–36	1,296
3	34	–24	576
4	50	–8	64
5	56	–2	4
6	58	0	0
7	60	2	4
8	66	8	64
9	82	24	576
10	94	36	1,296
11	96	38	1,444
Summary information			
	Central tendency		*Dispersion*
	Summation of ratings = 638		Summation of squared deviations = 6,768
			Average of squared deviations (variance) = 615.3
	$N = 11$		
	$\mu = 58$		$\sigma = 24.8$

in both scenarios, a mean rating of 58. However, in population A student ratings are more spread out—the distribution has a higher standard deviation—than in population B. In discussing population A, we will provide a step-by-step guide for arriving at the standard deviation. Populations A and B are unrealistic in two respects. First, both show calculations that have been performed on a population. Researchers perform calculations on samples, not populations. But here we bend reality so that we can introduce appropriate terminology and lay necessary groundwork. Second, both scenarios depict the student population as having 11 members, not the more realistic 20,000 students we have been using as an example. This is done to make the math easier to follow. After these simplifications have served their purposes, we will restore plausibility to the populations.

Table 6-1 presents student population A. As discussed earlier, a sample size is denoted by a lowercase *n*. In contrast, a population size is denoted by an uppercase *N*. In Table 6-1, then, $N = 11$. The thermometer ratings given by each member of the population are in the "Democratic rating" column, from 20 for the student with the coolest response to the Democrats to 96 for the most pro-Democratic student. To arrive at the mean rating for the population, we divide the summation of all ratings (20 + 22 + 34 + . . .), 638, by the population size, 11, which yields 58. Unlike sample statistics, which (as we will see) are represented by ordinary letters, population parameters are always symbolized by Greek letters. A population mean is symbolized by the Greek letter μ (pronounced "mew"). Thus, in Table 6-1, $\mu = 58$. This is a familiar measure of central tendency for interval-level variables.

How might we summarize variation in ratings among the students in this population? A rough-and-ready measure is provided by the **range,** defined as the maximum actual value minus the minimum actual value. So, in this example, the range would be the highest rating, 96, minus the lowest rating, 20, a range of 76. In gauging variation in interval-level variables, however, the measure of choice is the standard deviation. The standard deviation of a population is symbolized by the Greek letter σ ("sigma"). As its name implies, the standard deviation measures variation as a function of deviations from the mean of a distribution. The first step in finding the standard deviation, then, is to express each value as a deviation from the mean or, more specifically, to subtract the mean from each value.

Step 1. Calculate each value's deviation from the mean:

$$(\text{Individual value} - \mu) = \text{Deviation from the mean.}$$

A student whose rating is below the population mean will have a negative deviation, and a student who gave a rating that is above the mean will have a positive deviation. An individual with a rating equal to the population mean will have a deviation of 0. In Table 6-1, the deviations for each member of the population are shown in the column labeled "Deviation from the mean." These deviations tell us the locations of each population member relative to the population mean. So, for example, Student 1, who rated the Democrats at 20 on the scale, has a deviation of –38, 38 units below the population mean of 58. Student 7, who rated the Democratic Party at 60, is slightly above the population mean, scoring the Democrats 2 points higher than the population mean of 58. Deviations from the population mean provide the starting point for figuring out the standard deviation.

Step 2. Square each deviation. All measures of variation in interval-level variables, including the standard deviation, are based on the square of the deviations from the mean of the distribution. In Table 6-1, these calculations for each student in the population appear in the column labeled "Squared deviation from the mean." Squaring each individual deviation, of course, removes the minus signs on the negative deviations, those members of the population who gave ratings below the population mean. Notice, for example, that the square of Student 1's deviation, –38, is the same as the square of Students 11's deviation, 38. Both square to 1,444. Why perform a calculation that treats Student 1 and Student 11 as equal, when they clearly are not equal at all? Because, in the logic of the standard deviation, both of these students make equal contributions to the *variation* in ratings. Both lie an equal distance from the population mean of 58, and so both deviations figure equally in determining the dispersion of ratings around the mean.

Step 3. Sum the squared deviations. If we add all the squared deviations in the "Squared deviation from the mean" column, we arrive at the sum 6,768. The summation of the squared deviations, often called the *total sum of squares,* can be thought of as an overall summary of the variation in a distribution. When calculated on real-world data with many units of analysis, the total sum of squares is always a large and seemingly meaningless number. However, the summation of the squared deviations becomes important in its own right when we discuss correlation and regression analysis (see Chapter 8).

Step 4. Calculate the average of the sum of the squared deviations. The average of the squared deviations is known by a statistical name, the **variance.** The population variance is equal to the sum of the squared deviations divided by *N.* (*Special note:* To calculate the variance for a sample, you would divide the sum of the squared deviations by *n* – 1. This is discussed below.) For the population depicted in Table 6-1, the variance is the summation of the squared deviations (6,768) divided by the population size (*N* = 11), which yields an average of 615.3. Notice that, as with any mean, the variance is sensitive to values that lie far away from the mean. Students toward the tails of the distribution—Students 1 and 2 on the low end and Students 10 and 11 on the high end—make greater contributions to the variance than students who gave ratings that were closer to the population mean. That's the beauty of the variance. If a population's values cluster close to the mean, then the average of the squared deviations will record the closer clustering. As deviations from the mean increase, then the variance increases, too.

Table 6-2 Central Tendency and Variation in Democratic Thermometer Ratings: Hypothetical Scenario B

Student	Democratic rating	Deviation from the mean	Squared deviation from the mean
1	25	–33	1,089
2	34	–24	576
3	50	–8	64
4	55	–3	9
5	56	–2	4
6	58	0	0
7	60	2	4
8	61	3	9
9	66	8	64
10	82	24	576
11	91	33	1,089
Summary information	*Central tendency* Summation of ratings = 638		*Dispersion* Summation of squared deviations = 3,484 Average of squared deviations (variance) = 316.7
	N = 11 μ = 58		σ = 17.8

Step 5. Take the square root of the variance. The population parameter of current concern, the standard deviation, is based on the variance. In fact, the standard deviation is the square root of the variance. The standard deviation (σ) for the population of students in scenario A, then, is the square root of 615.3, or $\sqrt{615.3} = 24.8$.

Turn your attention to Table 6-2, which depicts a second possibility for the distribution of Democratic ratings in the student population. The mean Democratic rating for population B is the same as population A, μ = 58, but the scores are not as spread out. Evidence of this lower dispersion can be found in every column of Table 6-2. Notice that the range is equal to 91 – 25 = 66 (compared with 76 for population A) and that there are fewer double-digit deviations from the population mean. Most noticeable are the lower magnitudes of the squared deviations, which sum to 3,484 (compared with 6,678 for population A), and the variance, which is equal to 316.7, substantially less than the value we calculated for the more dispersed population (615.3). Taking the square root of the variance, we arrive at a σ equal to 17.8, which is 7 points lower than the standard deviation of population A (σ = 24.8). As we will demonstrate, a statistic computed on a random sample from population A will have a higher amount of random sampling error than will a statistic computed on a random sample drawn from population B.

n *and* σ

Let's pause and review the statistical components discussed thus far.

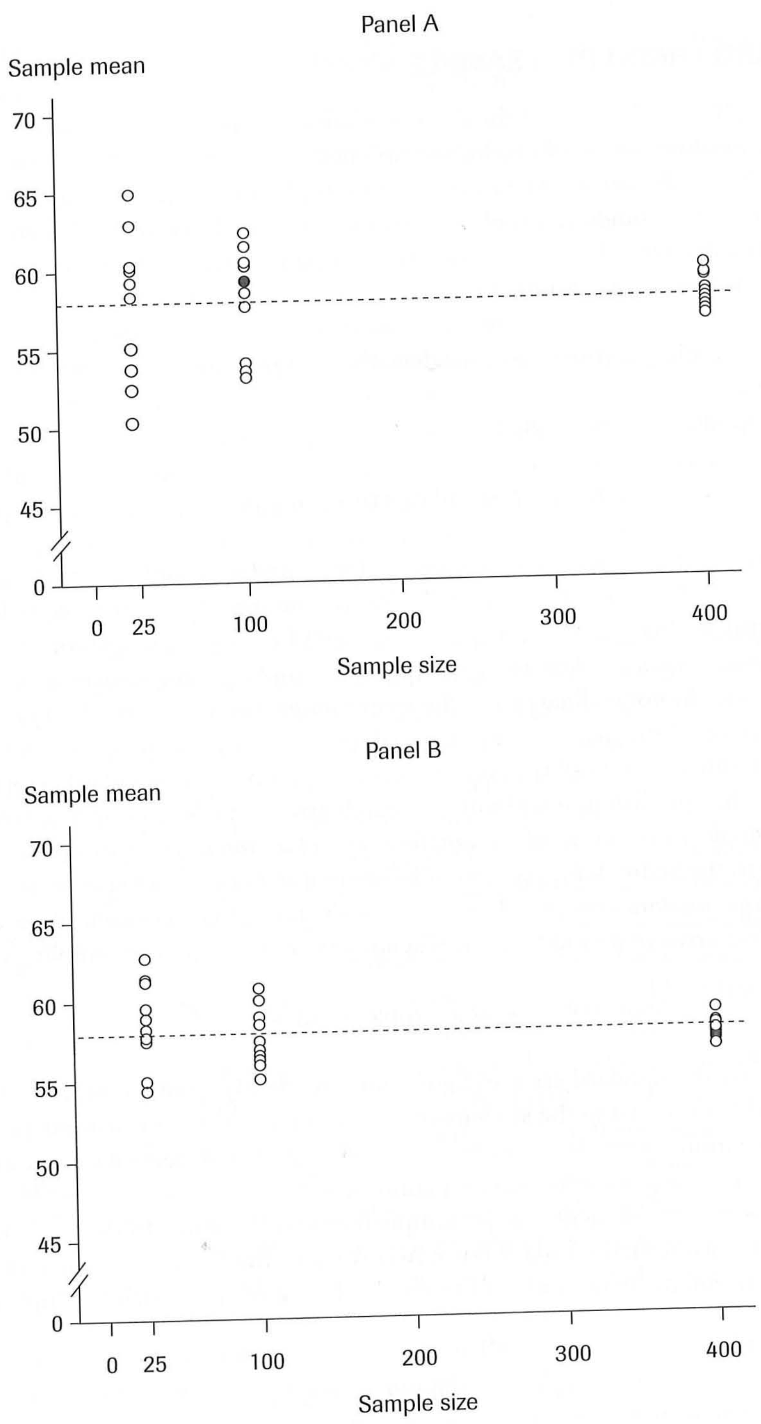
Sample size component: As the sample size goes up, random sampling error declines as a function of the square root of the sample size.
Variation component: As variation goes up, random sampling error increases in direct relation to the population's standard deviation.

Now, we will take a firsthand look at how these components work together. Again consider population A and population B—only this time think of them in a much more realistic light. Instead of a mere 11 members, each population now has 20,000 students. Just as its counterpart in Table 6-1, the distribution of the 20,000 students in population A has a mean equal to 58 and a standard deviation equal to 24.8. And, just as in Table 6-2, the distribution of the 20,000 students in population B has a mean equal to 58 and a standard deviation equal to 17.8. Having artificially created these realistic populations, we can ask the computer to draw random samples of different sizes from each population.[4] We can then calculate and record the mean Democratic rating obtained from each sample.

The results are presented in Figure 6-2. All the sample means displayed in panel A are based on the same student population—a population in which μ = 58 and σ = 24.8. All the sample means displayed in panel B were drawn from a student population in which μ = 58 and σ = 17.8. The dashed horizontal line in each panel shows the location of the true population mean, the parameter being estimated by the sample means. For each population, the computer drew ten random samples of $n = 25$, ten random samples of $n = 100$, and ten random samples of $n = 400$. So, by scanning from left to right within each panel, you can see the effect of sample size on random sampling error. By moving between panel A and panel B, you can see the effect of the standard deviation on random sampling error. (So that we don't lose track of our student researchers, their sample's mean of 59 appears as a solid dot in the $n = 100$ group in panel A. We return to this example below.)

Consider the set of sample means with the largest error component, the samples of $n = 25$ in panel A. Even though three or four of these sample means come fairly close to the population mean of 58, most are wide of the mark, ranging in value from the chilly (a mean Democratic rating of 50) to the balmy (a mean rating of 65). A small sample size, combined with a dispersed population parameter, equals a lot of random error. As we move across panel A to the ten sample means based on $n = 100$, we get a tighter grouping and less wildness, but even here the means range from about 53 to 62. The samples of $n = 400$ return much better precision. Four of the ten sample means hit the population mean almost exactly. Plainly enough, as sample size increases, error declines. By comparing panel A with panel B, we can see the effect of the population standard deviation on random sampling error. For example, notice that the ten samples of $n = 25$ in panel B generate sample statistics that are about as accurate as those produced by the samples of $n = 100$ in panel A. When less dispersion exists in the population parameter, a smaller sample can sometimes yield relatively accurate statistics. Naturally, just as in panel A, increases in sample size bring the true population mean into

Figure 6-2 Sample Means from Population with μ = 58 and σ = 24.8 (Panel A) and σ = 17.8 (Panel B)



Note: Hypothetical data. Hypothetical student group's sample mean is represented by the solid dot in the $n = 100$ group in panel A. Dashed horizontal line shows location of true population mean (μ = 58).

clearer focus. At $n = 400$ in panel B, six of the ten sample means are within a few tenths of a point of the true population mean. A larger sample, combined with lower dispersion, equals less random error and greater confidence in a sample statistic.

THE STANDARD ERROR OF