

Chapter 5

Populations and Samples: The Principle of Generalization

The remaining major component of the scientific method to be discussed is the process of scientific or statistical generalization. Generalization is a very common human process. We all draw conclusions about reality from a limited amount of experience. This saves us effort, but it can mislead us, because our experiences may be so limited or selective that the conclusions drawn from them are quite wrong. If I purchase an automobile which breaks down frequently, I might tell all my friends that “all Belchfires are pieces of junk”. But maybe only *my* Belchfire is a piece of junk, and all the others are very high quality. Another person may have a wonderful experience with a Belchfire and conclude that Belchfires are the finest automobiles made.

As anyone who has participated in a wine tasting will tell you, you pronounce a wine to be dry, oaky, acid, or fruity on the basis of a single sip. And when you make such a statement, you are not really talking only about that sip (a limited observation), but your pronouncements are really about the remaining untasted wine in the bottle (and perhaps, all the other bottles of the same type produced by that vineyard that year). In a similar vein we can look at dating. It starts with limited observation, (spending a few evenings at the movies, going to parties, perhaps studying together), which is followed by a decision to either continue seeing this person or to break off the relationship. The decision is based on (and reflects) a generalization from the limited past experiences with this person to what future and unobserved experiences with him or her would be like. And how many of you have ever watched the first five minutes or so of a situation comedy, only to turn it off, vowing never to tune in again?

Each of these examples relies on the judgment of a single individual, using whatever criteria the person deems appropriate. In contrast, science uses mathematics and statistics to standardize the generalization process, and provide some ways to describe the trustworthiness of a generalization. The same rules of observation and decision are used by different individuals, so the process does not depend on the whims of any individual. As we've mentioned before, scientific truth must be objective and capable of being shared among the whole scientific community. Statistical generalization provides this common ground for agreement.

Statistical generalization is the process where the scientist takes his or her conclusions, which are based on the observation of a limited number of cases, and extends these conclusions to cover all the other unobserved cases that fit in the same category. The set of limited observations is called the *sample* and the total number of cases is the *population* or *universe*.

Generalization is very much a part of all science, including communication science. Political polling organizations interview relatively small samples of eligible voters in order to find out what the political preferences of the voting public are, or how their candidate is faring. The A.C. Nielsen Company relies on observing the viewing behaviors of a small number of households to make statements about what the entire audience is viewing. Detroit auto manufacturers pull a small number of cars off the assembly line and test them to get a measure of the quality of all the cars produced during a particular time period.

Advantages and Disadvantages of Generalization

Why do we generalize? Probably the best reason we can give is that it is efficient. It saves us a lot of time, effort, and money. Look at the examples above in the light of efficiency. Perhaps political pollsters would really like to interview every registered voter and ask each of them how he or she feels about a particular candidate, and for whom he or she plans to vote in the next presidential election. However, the amount of time and money required to carry out such a task would be enormous. Furthermore, the information obtained would probably be useless by the time it was available. Suppose that a candidate for political office is interested in finding out whether a change in political advertising strategy is necessary in the remaining two weeks before an election. Spending a couple of days interviewing a sample of prospective voters will yield an answer quickly enough so that changes can be made before the election. It is highly doubtful that a canvass of all prospective voters could yield the needed information in time. Similarly, before we could ask all television households in the U.S. about the shows that they watch, the current television season would be long past.

However, there are also other reasons for generalizing to a large group from a smaller number of observations, such as the lack of an alternative. Suppose a manufacturer of light bulbs wants to know how long his bulbs really last. She chooses a selection of bulbs from a particular production run, and measures the number of hours each bulb lasts before it burns out. The average lifetime of the tested bulbs then will be ascribed to the remaining untested bulbs in the same production run. In this case the manufacturer generalizes because she has no alternative. If she had to test the entire set of products, she would have nothing left to sell. This illustrates another reason for generalizing from a limited sample to a full population: sometimes the process of observation is destructive of the actual cases being observed. Fortunately this is infrequently the case in communication research, although the research process can sometimes disrupt the very phenomenon that is being studied (see Chapter 13).

The above examples point out the advantages of studying only a limited number of cases and generalizing to the others. But there are also problems introduced by generalization.

One problem lies in generalizing from a limited set of observations to the larger set, when the limited set is not representative of the larger set. A sip of wine from a bottle which has been exposed to heat and sunlight would provide a taste experience which might not be representative of an entire production run. To interview only middle-class males about their family communication patterns will yield an inaccurate picture of the communication patterns of the entire population. To generalize from any sample which is not representative will usually result in an incorrect characterization of the population.

What do we mean by representative? Basically, that all the characteristics in a sample drawn

from a population are present in the same proportion as they are in the original population. If the population of the U.S. is 52% female, the proportion of females in a national sample should be approximately 52%. Likewise, if 10% of the population is below the poverty level, 10% of the sample should also fall in this region, and so forth for every other relevant characteristic of the population.

For the average person acting as a naive scientist, the problem of the representativeness of observations is often not a particularly pressing one. Also, it may be very difficult to determine whether a set of observations is indeed representative. However, to the scientist the problem of representativeness is a key problem. Consequently the scientific method contains a number of prescriptions for selecting cases for observation which helps assure representative samples.

Samples and Populations

As we've seen, a scientist formulates a theory about a phenomenon, and uses it to generate specific hypotheses. For instance, he may predict that personal income increases are associated with increases in newspaper readership, or that female gender is associated with greater sensitivity to nonverbal cues, or that increases in environmental uncertainty is associated with decreased centralization of communication in organizations. In formulating these predictions, the scientist is implicitly stating that the predictions made by the hypothesis will hold for all members of the population, not just for the sample being observed to test the hypothesis. That is, she is saying that they are general statements.

To be totally confident that any findings are really general statements about the population, we would have to inspect the entire population. We would have to say that all males were tested for nonverbal sensitivity, as were all females, and that the predicted difference indeed held up; that income and readership were measured in all newspaper subscribing households, etc. If we look at every case in the population, we can be completely confident that we've described the relationships correctly.

But this is usually impossible. It is obvious that a scientist will hardly ever be in a situation where the whole population can be scrutinized. We will never be able to contrast all the males in the U.S. to all the females, or, for that matter, to ask every registered voter who his or her favorite candidate for the presidency is. Since hypotheses generally cannot be tested on populations, the scientist will have to be content with selecting a subset from the population. The hypotheses can then be tested on this subset from the population and the results of the test generalized to the population.

Such a subset is the sample. It is considered to be representative of the population. Notice we did not say that the sample actually is representative of the population, merely that it is *thought* to be so. It must be selected in such a way that any conclusions drawn from studying the sample can be generalized.

The scientist is fully aware that generalizations to the population will be made, and that is why it is so critical that the scientist follow certain rules to assure that the sample on which the test was carried out can be considered to be representative of the population. These rules have been devised to prevent biases from creeping in. The introduction of any biases would make samples unrepresentative, and the conclusions not generalizable.

Sample Bias

There are two major sources of sample bias: *selection bias* and *response bias*. Selection bias is introduced by the actions of the researcher; response bias is introduced by the actions of the units (such as individuals) being sampled.

Selection bias. This type of bias is introduced when some potential observations from the population are excluded from the sample by some systematic means. For example, if we were to carry out research on political information seeking by conducting in-house interviews, we might be tempted to systematically avoid those houses with "Insured by Smith & Wesson" signs, houses at the end of long muddy driveways, apartments on the fourth floor of rickety tenement buildings, and all the haunted-looking houses on Elm Street. By avoiding these places we would be selecting nice, reliable-looking houses, probably inhabited by nice, reliable people (like us, incidentally). As nice as this sample might be, it's probably not very representative of the population of voting citizens. If we try to generalize the conclusions of hypothesis tests from this sample to the whole popu-

lation, we're probably going to be wrong.

To be a bit formal, a sample contains selection bias whenever each unit in the population does not have an equal probability of being chosen for the sample. The idea of having an equal probability of selection for all units in the population is called the *equal likelihood principle*.

Response Bias

Response bias is difficult to avoid. It occurs when the cases chosen for inclusion in the sample systematically exclude themselves from participation in the research. For instance, if we are interested in studying the relationship between the kinds of job rules in an organization and certain aspects of subordinate-superior communication, we will need to observe a large number of organizations. But it is quite likely that those organizations which consent to having employees interviewed are different in many respects (other than cooperativeness) from those organizations which decline to participate. These differences may affect the hypothesis being tested. After all, those companies who consented to have interviewers come in may have done so because they have no serious organizational problems. If we assume that other companies may have turned us down because they have problems which they don't want to air, then we see the potential for error. In our biased sample we may find that the imposition of restrictive job rules does not affect the amount of superior-subordinate communication. But in the organizations which we have not observed, more job rules may lead to decreased communication. If we generalize to all organizations from our biased sample, we will simply be drawing the wrong conclusions.

Randomization

Sample bias can result from either the action of the scientist (selection bias) or from the units of the sample (response bias). The scientist must take steps to avoid both types of bias, so that the results of his study can be safely generalized to the population. The rules that the scientist must follow in drawing samples are designed to counteract or defuse the often-unconscious tendencies to select or overlook certain observations in favor of others (selection biases), or to detect and deliberately compensate for response biases.

The method of avoiding selection bias is quite simple: base the selection process on pure chance. This process is called *random sampling*. Randomness can be defined as the inability to predict outcomes. If we select street addresses by pulling street names and house numbers out of a hat containing all street names and numbers, we'll not be able to predict which houses will be selected for our sample. Further, each house will have exactly the same probability of being selected, which satisfies the Equal Likelihood Principle. It then follows that we will not be able to predict whether any person selected in our political information seeking study is a nice, middle-class person or is a member of some other socioeconomic class. All socioeconomic classes should therefore be represented in the sample in their correct proportions, rather than being systematically excluded. The final sample should then have non-middle class observations proportional to the number of non-middle class persons in the population, and we can safely generalize our findings to the whole population.

Unfortunately, random selection will not help us avoid response bias. We can select the sample randomly, but if respondents eliminate themselves, the result is a biased sample. But there are a number of techniques to detect response bias and to compensate somewhat for it. In general, sample bias is estimated by comparing known characteristics of the population, such as age and income levels obtained from the U.S. census, to the same variables measured in the sample. If these values are highly discrepant, then the sample is probably biased. This bias can be partially overcome by sampling techniques like over sampling (deliberately selecting additional cases similar in the known characteristics to those who refused to respond) and weighting (treating some observations as more important than others). These techniques violate the Equal Likelihood Principle, but they do so in a way that partially compensates for known sample biases. In essence, they introduce a canceling bias that produces a final sample that is closer to one which adheres to the Equal Likelihood Principle. Over sampling and weighting are covered in more detail in the next chapter.

Sampling Error

Even an unbiased sample can be less than a completely accurate representation of the population. The degree to which the value of an estimate calculated from the sample deviates from the actual population value is called *sampling error*. If we select observations for a sample randomly, we still might get an unrepresentative sample by simple bad luck, just as we might draw 10 consecutive losing hands at blackjack. It's improbable, but it can happen. For example, even if we choose house addresses randomly, we might get all middle-class households. This is improbable, but it's possible.

It is important to understand that the error in an unbiased sample is qualitatively different from the error in a biased sample. Sampling error in an unbiased sample is not systematic, and it becomes less and less with each additional observation added to the sample, as it's less and less likely that we'll continue to pick improbable units, and more and more likely that we'll pick units which cancel out the error in our previous observations (we'll talk more about this below). If we continue to add house addresses to our political information sample, it is increasingly unlikely that the sample will include only middle-class families.

Sampling error is not systematic, that is, it is random. This means that we are just as likely to make estimates which are too high as we are to make estimates which are too low. This makes sampling error different from sample bias, where the estimates will always be systematically too high or too low. If we draw a number of biased samples from a population, we will always tend to get an estimate that is systematically too high or too low. But if we draw a number of unbiased samples, the sample estimates will be about equally spread between those which are too high and those which are too low.

An Illustration of Sampling Error

One of the best ways of illustrating the concept of sampling error is through a simple numerical example.

To keep the example simple, let us assume that we have a population with only five persons, all mothers of schoolaged children. We are going to use samples to estimate the number of conversations about schoolwork these mothers have had, on the average, with their children in the past week, and generalize this sample estimate to the entire population. The actual number of conversations each person has had in the past week is shown in Table 5-1.

One of the ways in which we can describe the communication about schoolwork that occurs in this population is to compute the average of these five scores to provide us with an single index that characterizes the population as a whole, without enumerating all the individual cases as in Table 5-1. We'll call this average *M*, short for Mean. Its value is:

Table 5-1 Actual Number of Conversations Between Mother and Child

Mother A: 5
Mother B: 6
Mother C: 7
Mother D: 8
Mother E: 9

$$M = (5 + 6 + 7 + 8 + 9) / 5 = 7.00 \text{ (5.1)}$$

The “average” member of the population talks about schoolwork with her child 7 times per week.

Now let us take all the possible samples of two individuals ($N = 2$) from this population. We’ll do this in a special way called “sampling with replacement”. To do this, we randomly select the unit to measure, then we replace that unit in the population before we select the next unit. Why would we do this? Because each unit must have an identical probability of being selected for the sample—the Equal Likelihood Principle. Since we have five cases in the population, each case has a chance of selection of one out of five, or 20%.

If we select one case and don’t replace it in the population, the chance of any of the remaining four being selected into the sample is now one out of four, or 25%. This violates the Equal Likelihood Principle of random sampling, because the probabilities of being selected into the sample are not the same for all the units in the population. If we return any unit to the population before choosing the next unit to observe, the odds of any unit (including the one chosen in the first observation) being chosen for the second observation in the sample then remains at one out of five, or 20%.

This kind of sampling is almost never done in communication research. Normally our population is so large that the probability differences between sampling with replacement and without replacement (where any unit chosen in the sample is not eligible for later inclusion in the sample) are trivial. For example, if our population is 10,000, the first observation has a probability of being chosen of $1 / 10,000$ or .0001000. The probability of any of the remaining units being chosen next is $1 / 9999$ or .00010001, which is very close to the same probability as the first observation. Such small violations of equal likelihood do not bias the sample to any important extent. But with a small population, like the one in this example, we have to sample with replacement, or the sample will be seriously biased.

Choosing all possible samples of two observations from the population of five gives us the values for Mean Conversations Per Week shown in Table 5-2. Notice that in this table we use the following symbol to represent the mean of the variable for a sample:

$$\bar{X}$$

Notice that our sampling procedure generates 25 different samples, but it generates fewer than 25 different means. If you look closely at Table 5-2, you’ll see that only nine different values of the mean were observed. These means and the number of times that each was generated are presented in Table 5-3. This table is important. If we select observations randomly, then we are equally likely to see any one of the pairs of values in Table 5-2. But five of the 25 combinations (20%) have a mean of 7.0, 16% have means of 6.5 and 7.5, etc. That means that if we draw any sample of $N=2$ from the population, the most probable value which we are likely to get from the sample is 7.0, which is

Table 5-2 All Possible Samples of $N = 2$ and Their Associated Means

Sample Values	Mean	Sample Values	Mean	Sample Values	Mean	Sample Values	Mean	Sample Values	Mean
5, 5	5.0	6, 5	5.5	7, 5	6.0	8, 5	6.5	9, 5	7.0
5, 6	5.5	6, 6	6.0	7, 6	6.5	8, 6	7.0	9, 6	7.5
5, 7	6.0	6, 7	6.5	7, 7	7.0	8, 7	7.5	9, 7	8.0
5, 8	6.5	6, 8	7.0	7, 8	7.5	8, 8	8.0	9, 8	8.5
5, 9	7.0	6, 9	7.5	7, 9	8.0	8, 9	8.5	9, 9	9.0
Sampling with replacement from a population consisting of the values 5, 6, 7, 8 and 9 (See Table 5-1).									

Table 5-3 Means and Their Frequency of Observation for All Possible Samples of $N = 2$

	<i>Mean</i>	<i>Frequency</i>	<i>Proportion</i>
	5.0	1	$1/25 = .04$
	5.5	2	$2/25 = .08$
	6.0	3	$3/25 = .12$
	6.5	4	$4/25 = .16$
True Population Mean →	7.0	5	$5/25 = .20$
	7.5	4	$4/25 = .16$
	8.0	3	$3/25 = .12$
	8.5	2	$2/25 = .08$
	9.0	1	$1/25 = .04$
		25	1.00
Values taken from a population consisting of the values 5, 6, 7, 8, and 9 (See Tables 5-1 and 5-2)			

exactly the population mean! That's just what we want from a sample, since if we generalize that value to the population, we'll be exactly right. Furthermore, if we get (by chance) one of the next most probable values (6.5 or 7.5), we'll be in error by only .5 conversations per week. The further the estimated sample value is from the true population value (in other words, the more it is in error), the less likely we are to observe it when we randomly select a sample.

Put another way, small under- or over-estimates occur more frequently than large under- or over-estimates. That is, when we sample randomly we are more likely to obtain samples which are representative of the population than we are to get samples which are less representative. The odds are with us. But samples which are not representative of the population (ones with substantial sampling error) nonetheless *do* occur, and we can't discount them totally. There are other important patterns in Tables 5-2 and 5-3.

First of all, the observed sample means are just as likely to over-estimate the true population mean as they are to under-estimate it. Sample means of 5.0, 5.5, 6.0 and 6.5 all under-estimate the population mean; in the aggregate they were found for 10 combinations of two observations (Table 5-2). The sample mean values 7.5, 8.0, 8.5 and 9.0 all over-estimate the population mean of 7.0; they, too, occur 10 times in total.

Second, over-estimates and under-estimates of the same magnitude occur with equal frequency. Consider the sample mean of 5.0, and the under-estimate of $(7.0 - 5.0) = 2.0$. This sample mean occurs only once in the 25 combinations, so it will be seen only 4% of the time in a random sample of $N = 2$. A sample mean which over-estimates by the same amount (the value 9.0, 2.0 conversations per week above the true population value) was also observed only once. Similarly, over-estimates of .5 (means of 7.5) and under-estimates of the same size (means of 6.5) each will occur 4 times or in 16% of the random samples; under- and over-estimates of 1.0 will each occur 12% of the time, etc.

The third observation, which follows naturally from the first two, is that all the over-estimates and underestimates cancel one another, so that across a large number of samples we can obtain an accurate estimate of the population value of any variable. If we keep sampling, we'll tend to get a better picture of the population.

The ideas contained in this simple example are fundamental, and can be expressed in very elegant mathematical proofs. But for our purposes, it's enough to remember these simple principles: although there is error associated with the process of random sampling, the *most likely* sample value which we will obtain is the correct population value. And the more observations we make, the

more likely we are to arrive at (or near) this value. The value obtained from a random sample is unbiased, as one cannot predict whether error is likely to be above or below the population value, and so we will not systematically under- or over-estimate the true population value. We can thus use the sample value to generalize to the population in an unbiased way.

But the above example assumes that we have not introduced any sample bias during our sampling procedures. Sample biases produce systematic errors in our estimates. The following example shows what happens when the sample is biased.

An Example of Sample Bias

We will carry out the same procedures as we did in the previous example, but we will introduce a rather severe form of bias in sampling: we will give the value 9 no chance of being selected into the samples of $N = 2$ that will be drawn from the population. Perhaps those who spend large amounts of time discussing schoolwork with their children are too busy to cooperate with our interviewer, so response bias is introduced. Or perhaps our interviewer has done all his interviewing with professional women who have less time to interact with children, thus introducing a selection bias. In either case, the result is the same: the values 5, 6, 7, and 8 will each have a 1 in 4 chance of being included in the sample (25%), and the value 9 will have no probability of being included (0%).

Table 5-4 contains all the different samples and sample means we obtain when we sample with this bias. This table differs from the one presented in Table 5-2 only by the elimination of any sample which contains the value 9.

Only sixteen different combinations of values are observed now. The resulting distribution of sample means and the frequency with which they are obtained is shown in Table 5-5.

If we contrast the pattern of sample means we observed while sampling with a bias (Table 5-4)

Table 5-4 All Possible Samples of $N = 2$ and Their Associated Means

Sample Values	Mean	Sample Values	Mean	Sample Values	Mean	Sample Values	Mean
5, 5	5.0	6, 5	5.5	7, 5	6.0	8, 5	6.5
5, 6	5.5	6, 6	6.0	7, 6	6.5	8, 6	7.0
5, 7	6.0	6, 7	6.5	7, 7	7.0	8, 7	7.5
5, 8	6.5	6, 8	7.0	7, 8	7.5	8, 8	8.0
Values and their associated means computed when sampling with replacement from a population consisting of the values 5, 6, 7, 8, and 9 with the value 9 (# of conversations for Mother E) excluded from the samples.							

Table 5-5 Means and Their Frequency of Observation from a Biased Sample of $N = 2$

	Mean	Frequency	Proportion
	5.0	1	1/16 = .0625
	5.5	2	2/16 = .1250
	6.0	3	3/16 = .1875
	6.5	4	4/16 = .2500
True Population Mean →	7.0	3	3/16 = .1875
	7.5	2	2/16 = .1250
	8.0	1	1/16 = .0625
		16	1.0000
Values taken from a population consisting of the values 5, 6, 7, 8, and 9 (See Tables 5-1 and 5-2)			

to the pattern of those found in an unbiased sample (Table 5-2), an important difference emerges.

The most probable value for the mean which we will likely find using such a biased sample is 6.5, and this is *not* the true population mean of 7.0. The observed sample means are also less likely to over-estimate than they are to underestimate the true population mean. Sample means of 5.0, 5.5, 6.0 and 6.5 which under-estimate the population mean are observed in 10 of the 16 samples. The only sample mean values that over-estimate the population mean of 7.0 are 7.5 and 8.0, and together they occur only three times. For example, consider the sample mean of 5.0, which represents an under-estimate of $(7.0 - 5.0) = 2.0$. This sample mean was observed once. A sample mean which would have overestimated by the same amount (a sample mean of 9.0) was never observed. Over-estimates and under-estimates of the true population are not equal, and they do not cancel one another, even if a large number of samples are drawn. When we have a biased sample in which all the units in a population do not have an equal chance of being selected into a sample, over-estimates will fail to cancel underestimates, and we will either consistently over-estimate or consistently under-estimate the population characteristic being observed. Figure 5-1 shows this bias graphically.

This simple example illustrates the danger of generalizing from biased samples. If we generalize from this sample, we will probably make an incorrect statement about the number of conversations that the “average” mother has with her child. And drawing more samples will not help, as we will, for this example anyway, consistently under-estimate the true population value. As long as bias is present, we will get incorrect results.

Sample Size

The last example in this chapter shows the effect of collecting more observations in a random sample. We’ve seen that a random sample without bias will tend to give the correct population value, but that most samples will probably contain some error in estimating this value. In Table 5-3, for example, we see that 80% of the samples will have some error in their estimate of the true popu-

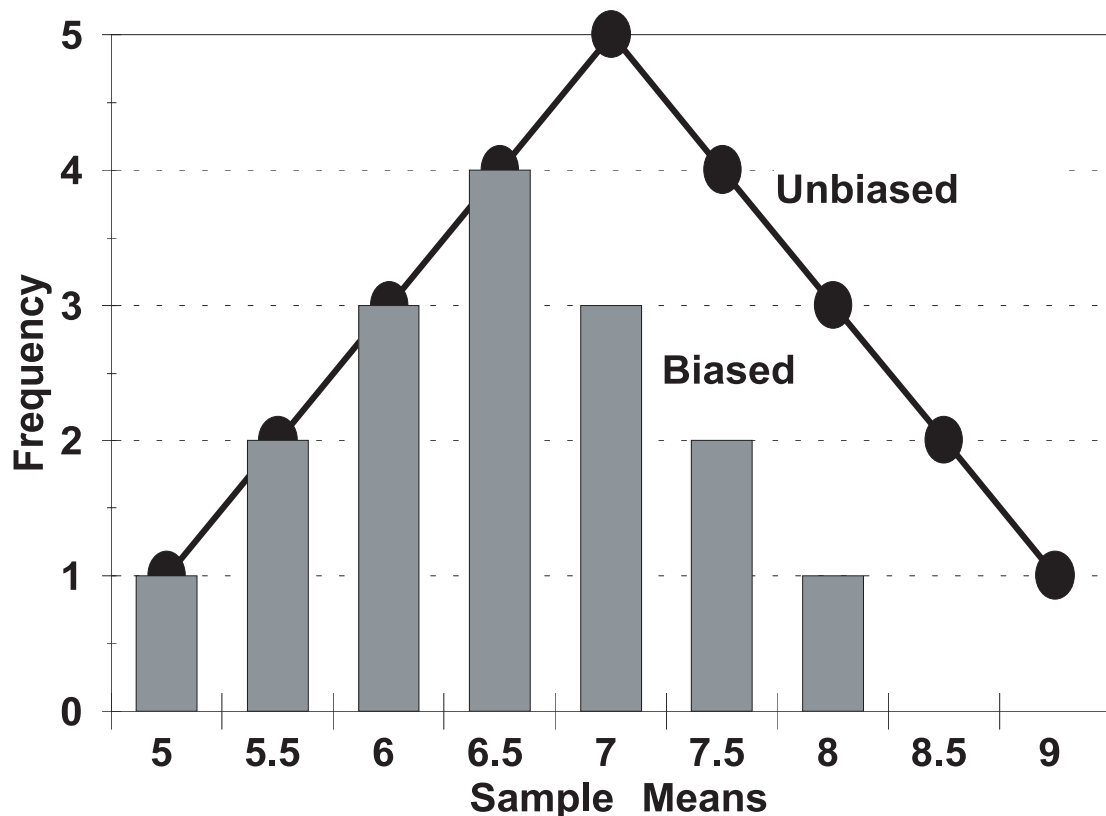


FIGURE 5-1 Distributions of Means from Biased and Unbiased Random Sampling with All Possible $N=2$ Samples

lation value. Since the name of the game is to obtain a sample which has the least possible discrepancy between sample characteristics and population characteristics, we want to see if we can improve the accuracy of a sample by increasing the number of observations. We will carry out the same sampling process as before, except that now we will select unbiased samples of three observations ($N = 3$), rather than $N = 2$ as before.

If we look at Table 5-6, one of the consequences of increasing sample size is rather dramatically obvious: the number of different samples that can be obtained when we take samples of $N = 3$ has risen to 125, in contrast to the 25 different samples that are possible when $N = 2$. But only 13 different sample means are obtained from these 125 samples. These sample means and the frequency with which they were observed are presented in Table 5-7.

To compare the accuracy of these samples of $N = 3$ with the $N = 2$ samples, we'll look at the percentage of means which come relatively close to the actual population value. For this example, let's look at means which fall plus or minus 1.0 from the population mean of 7.0. The sample means

Table 5-6 All Possible Samples of $N = 3$ and Their Associated Means

Sample Values	Mean	Sample Values	Mean	Sample Values	Mean	Sample Values	Mean	Sample Values	Mean
5,5,5	5.00	5,6,5	5.33	5,7,5	5.66	5,8,5	6.00	5,9,5	6.33
5,5,6	5.33	5,6,6	5.66	5,7,6	6.00	5,8,6	6.33	5,9,6	6.66
5,5,7	5.66	5,6,7	6.00	5,7,7	6.33	5,8,7	6.66	5,9,7	7.00
5,5,8	6.00	5,6,8	6.33	5,7,8	6.66	5,8,8	7.00	5,9,8	7.33
5,5,9	6.33	5,6,9	6.66	5,7,9	7.00	5,8,9	7.33	5,9,9	7.66
6,5,5	5.33	6,6,5	5.66	6,7,5	6.00	6,8,5	6.33	6,9,5	6.66
6,5,6	5.66	6,6,6	6.00	6,7,6	6.33	6,8,6	6.66	6,9,6	7.00
6,5,7	6.00	6,6,7	6.33	6,7,7	6.66	6,8,7	7.00	6,9,7	7.33
6,5,8	6.33	6,6,8	6.66	6,7,8	7.00	6,8,8	7.33	6,9,8	7.66
6,5,9	6.66	6,6,9	7.00	6,7,9	7.33	6,8,9	7.66	6,9,9	8.00
7,5,5	5.66	7,6,5	6.00	7,7,5	6.33	7,8,5	6.66	7,9,5	7.00
7,5,6	6.00	7,6,6	6.33	7,7,6	6.66	7,8,6	7.00	7,9,6	7.33
7,5,7	6.33	7,6,7	6.66	7,7,7	7.00	7,8,7	7.33	7,9,7	7.66
7,5,8	6.66	7,6,8	7.00	7,7,8	7.33	7,8,8	7.66	7,9,8	8.00
7,5,9	7.00	7,6,9	7.33	7,7,9	7.66	7,8,9	8.00	7,9,9	8.33
8,5,5	6.00	8,6,5	6.33	8,7,5	6.66	8,8,5	7.00	8,9,5	7.33
8,5,6	6.33	8,6,6	6.66	8,7,6	7.00	8,8,6	7.33	8,9,6	7.66
8,5,7	6.66	8,6,7	7.00	8,7,7	7.33	8,8,7	7.66	8,9,7	8.00
8,5,8	7.00	8,6,8	7.33	8,7,8	7.66	8,8,8	8.00	8,9,8	8.33
8,5,9	7.33	8,6,9	7.66	8,7,9	8.00	8,8,9	8.33	8,9,9	8.66
9,5,5	6.33	9,6,5	6.66	9,7,5	7.00	9,8,5	7.33	9,9,5	7.66
9,5,6	6.66	9,6,6	7.00	9,7,6	7.33	9,8,6	7.66	9,9,6	8.00
9,5,7	7.00	9,6,7	7.33	9,7,7	7.66	9,8,7	8.00	9,9,7	8.33
9,5,8	7.33	9,6,8	7.66	9,7,8	8.00	9,8,8	8.33	9,9,8	8.66
9,5,9	7.66	9,6,9	8.00	9,7,9	8.33	9,8,9	8.66	9,9,9	9.00
Sampling with replacement from a population consisting of the values 5, 6, 7, 8, and 9.									

Table 5-7 Means and Their Frequency of Observation for All Possible Samples of $N = 3$

	<i>Mean</i>	<i>Frequency</i>	<i>Proportion</i>
	5.0	1	1/125 = .008
	5.33	3	3/125 = .024
	5.66	6	6/125 = .048
	6.00	10	10/125 = .08
	6.33	15	15/125 = .12
	6.66	18	18/125 = .144
True Population Mean →	7.00	19	19/125 = .152
	7.33	18	18/125 = .144
	7.66	15	15/125 = .12
	8.00	10	10/125 = .08
	8.33	6	6/125 = .048
	8.66	3	3/125 = .024
	9.00	1	1/125 = .008
		125	1.000
Values taken from a population consisting of the values 5, 6, 7, 8, and 9 (See Tables 5-1 and 5-2)			

that fall within this error bracket are, relatively speaking, close to the true population mean and reflect small amounts of sampling error. Those outside the bracket reflect more sampling error. The greater the proportion of means inside the bracket, the lesser the degree of sampling error.

As Table 5-8 shows, the proportion of sample means that fall within 1.0 of the true population mean is 76% when $N = 2$, and 84% when $N = 3$. Another way of stating this finding is this: If we were to take a sample of 2 observations from this population, we would have a 76% chance of getting an estimate that is within 1.0 of the true population value. If we increase the sample size to 3, the odds of observing a discrepancy of 1.0 or less improves to 84%. Looking at it negatively, when $N = 2$, we have a 24% chance of having sampling error greater than 1.0; when $N = 3$ this chance is reduced to 16%.

The general conclusion to be drawn from this exercise is that increased sample size is associated with decreases in sampling error. An increased sample size increases the representativeness of samples and makes generalization safer. Unfortunately, as we will see later, the relationship between sampling error and sample size is not a simple proportional one. As we increase the sample size more and more, we will always get improved accuracy, but the reductions in sampling error get smaller and smaller with each added observation. There are diminishing returns associated with adding observations to the sample, so there comes a point where it is not cost effective to add any more observations to a sample. We will also discuss way to determine the appropriate sample sizes in Chapter 12.

Summary

The process of statistical generalization is an integral part of the scientific process. From a limited number of observations, called the *sample*, we extend conclusions to a whole population, which contains all the units which we could observe. We sample and generalize for reasons of cost efficiency, speed in obtaining information, or because our observations interfere with the phenomenon being studied.

Table 5-8 Effect of Sample Size on Sampling Error

N=2		N=3	
Means within [+or-] 1.0 of Population Mean	Proportion	Means within [+or-] 1.0 of Population Mean	Proportion
6.00	.12	6.00	.080
6.50	.16	6.33	.120
7.00	.20	6.66	.144
7.50	.16	7.00	.152
8.00	.12	7.33	.144
Total:	.76	7.66	.120
		8.00	.080
		Total:	.84

Statistical generalization requires that samples be drawn in an unbiased fashion, or our conclusions will be in error. There are two types of bias against which the communication scientist must guard: *selection biases* in which each member of the population does not have an equal probability of being chosen for a sample; and *response biases* in which the member of the population chosen does not participate in the research. Both biases violate the Equal Likelihood Principle, which requires that all members of the population have an identical probability of being included in the sample. To meet the *equal likelihood principle*, some random procedure for selecting samples is typically used.

Even with pure random selection, there is still *sampling error* to contend with. By chance alone, we may select an unrepresentative sample from the population. However, the probability of selecting unrepresentative samples goes down as the size of the sample is increased. Larger samples are less likely to contain large sampling errors than are smaller samples, as the error introduced by new observations will tend to cancel previously introduced error. However, larger sample sizes will not correct sample bias. Biased samples will give systematically incorrect results.

Only when we are confident that whatever we observe in the sample is also true for the population can we make reasonable generalizations. For this reason, the researcher must guard against sample biases, and include as many observations in the sample as is practically reasonable.

References and Additional Readings

- Cochran, W.G. (1983). *Planning and analysis of observational studies*. New York: Wiley. (Chapter 2, "Statistical Induction").
- Hogben, L. (1970). Statistical prudence and statistical inference. In D.E. Morrison & R. Henkel (Eds.) *The significance test controversy: A reader* (pp. 22-56). Chicago: Aldine.
- Loether, H.J. and McTavish, D.G. (1974). *Inferential statistics for sociologists: An introduction*. Boston: Allyn & Bacon. (Chapter 1, "Descriptive and Inferential Statistics").