Lecture Notes

# Chapter 6: Sampling and Sampling Distributions

## Learning Objectives

1. Describe the aims of sampling and basic principles of probability.
2. Explain the relationship between a sample and a population.
3. Identify and apply different sampling designs.
4. Apply the concept of the sampling distribution.
5. Describe the central limit theorem.

## Chapter Outline

1. Introduction
	1. People are rarely able to study or observe everyone or everything we are interested in.
	2. Although prior chapters discussed various methods to analyze observations, these observations represent a fraction of all the possible observations.
	3. The primary problem in research situations is that there is too much information and not enough resources to collect and analyze it.
2. Aims of Sampling
	1. Researchers in the social sciences rarely have enough time or money to collect information about the entire group that interests them.
	2. Known as the **population**, this group includes all the cases (individuals, groups, or objects) in which the researcher is interested.
	3. We can learn a lot about a population if we carefully select a subset of it. This subset is called a **sample**.
	4. Through the process of **sampling**, selecting a subset of observations from the population, we attempt to generalize the characteristics of the larger group (population) based on what we learn from the smaller group (the sample).
	5. This is the basis of inferential statistics, that is, making predictions or inferences about a population from observations based on a sample. Thus, it is important how we select our sample.
	6. The term **parameter**, associated with the population, refers to measures used to describe the population we are interested in.
	7. When used to describe the population distribution, measures such as a proportion, a mean, or a standard deviation are referred to as parameters. Thus, a population mean, a population proportion, and a population standard deviation are all parameters.
	8. The term **statistic** refers to a corresponding characteristic calculated for the sample. A sample mean, a sample proportion, and a sample standard deviation are all statistics.
3. Basic Probability Principles
	1. We all use the concept of probability in everyday conversation. In the study of statistics, probability has a far more precise meaning.
	2. A variety of techniques are adopted by social scientists to select samples from populations. The techniques follow a general approach called probability sampling.
	3. We will briefly review some theories and principles of probability.
		1. A **probability** is a quantitative measure that a particular event will occur.
		2. It is expressed as a ratio of the number of times an event will occur relative to the set of all possible and equally likely outcomes.
		3. Probability is represented by a lower case *p*.
		4. *p* = Number of times an event will occur/Total number of events
		5. Probabilities range in value from 0 (the event will not occur) to 1 (the event will certainly occur).
		6. Probabilities can be expressed as proportions or percentages.
	4. Sometimes we use information from past events to help us predict the likelihood of future events. Such a method is called the relative frequency method.
		1. The observed relative frequencies are just an approximation of the true probability of occurrence of an observation.
		2. The true probabilities can only be determined if we were to repeat the study many times under the same conditions.
		3. Then, our long-run relative frequency (or probability) will approximate the true probability.
4. Probability Sampling
	1. Only one general approach, probability sampling, allows the researcher to use the principles of statistical inference to generalize from the sample to the population.
	2. **Probability sampling** is a method that enables the researcher to specify for each case in the population the probability of its inclusion in the sample.
		1. The purpose of probability sampling is to select a sample that is as representative as possible of the population.
		2. The sample is selected in such a way as to allow the use of the principles of probability to evaluate the generalizations made from the sample to the population.
		3. A probability sample design enables the researcher to estimate the extent to which the findings based on one sample are likely to differ from what would be found by studying the entire population.
	3. Although accurate estimates of sampling error can be made only from probability samples, social scientists often use nonprobability samples because they are more convenient and cheaper to collect.
		1. Nonprobability samples are useful under many circumstances for a variety of research purposes.
		2. Their main limitation is that they do not allow the use of the method of inferential statistics to generalize from the sample to the population.
	4. We will learn about three sampling designs that follow the principles of probability sampling:
		1. The simple random sample
		2. The systematic random sample
		3. The stratified random sample.
	5. The Simple Random Sample
		1. The simple random sample is the most basic probability sampling design, and it is incorporated into even more elaborate probability sampling designs.
		2. A **simple random sample** is a sample design chosen in such a way as to ensure that
			1. Every member of the population has an equal chance of being chosen.
			2. Every combination of *N* members has an equal chance of being chosen.
		3. Researchers usually use computer programs or tables of random numbers in selecting random samples. An abridged table of random numbers is reproduced in Appendix A. To use a random number table,
5. List each member of the population and assign the member a number.
6. Begin anywhere on the table and read each digit that appears in the table in order, up, down, or sideways; the direction does not matter, as long as it follows a consistent path.
7. Whenever we come across a digit in the table of random digits that corresponds to the number of a member in the population of interest, that member is selected for the sample.
8. Continue this process until the desired sample size is reached.
	1. The Systematic Random Sample
		1. This sampling method that is easier to implement than a simple random sample.
		2. The systematic random sample, although not a true probability sample, provides results very similar to those obtained with a simple random sample.

It uses a ratio, *K*, obtained by dividing the population size by the desired sample size: $ $

* + 1. **Systematic random sampling** is a method of sampling in which every *K*th member in the total population is chosen for inclusion in the sample after the first member of the sample is selected at random from among the first *K*th members in the population.
	1. The Stratified Random Sample
		1. A third type of probability sampling, the **stratified random sample** can be obtainedby
			1. Dividing the population into subgroups based on one or more variables central to our analysis and then
			2. Drawing a simple random sample from each of the subgroups.
		2. The choice of subgroups is based on what variables are known and what variables are of interest to us.
		3. In a **proportionate stratified sample,** the size of the sample selected from each subgroup is proportional to the size of that subgroup in the entire population.
1. Proportional sampling ensures the representation of the subgroup variable.
2. Proportionate sampling can result in the sample having too few members from a small subgroup to yield reliable information about them.
	* 1. In a **disproportionate stratified sample**, the size of the sample selected from each subgroup is deliberately made disproportional to the size of that subgroup in the population.
3. In such a sampling design, although the sampling probabilities for each population member are not equal (they vary between groups), they are known, and therefore, we can make accurate estimates of error in the inference process.
4. Disproportionate stratified sampling is especially useful when we want to compare subgroups with each other, and when the size of some of the subgroups in the population is relatively small.
5. The Concept of the Sampling Distribution
	1. One of the most important concepts in statistical inference is sampling distribution. The sampling distribution helps estimate the likelihood of our sample statistics, and therefore enables us to generalize from the sample to the population.
	2. The Population
		1. Let’s consider as our population the 20 individuals.
		2. Our variable, *Y*, is the income (in dollars) of these 20 individuals.
		3. The parameter we are trying to estimate is the mean income.
		4. The symbol µ represents population mean.
		5. The symbol σ represents the population’s standard deviation.
		6. Draw one sample, compute the mean--the statistic--for that sample and use it to estimate the population mean--the parameter.
	3. The Sample
		1. **Sampling error** is the discrepancy between a sample estimate of a population parameter and the real population parameter.
		2. By comparing the sample statistic with the population parameter, we can determine the sampling error.
	4. The Dilemma
		1. Although comparing the sample estimates of the average income with the actual population average is a perfect way to evaluate the accuracy of our estimate, in practice, we rarely have information about the actual population parameter.
		2. This, then, is our dilemma:
			1. If sample estimates vary and if most estimates result in some sort of sampling error, how much confidence can we place in the estimate?
			2. On what basis can we infer from the sample to the population?
	5. The Sampling Distribution
		1. The **sampling distribution** is a theoretical probability distribution of all possible sample values for the statistic in which we are interested.
		2. If we were to draw all possible random samples of the same size from our population of interest, compute the statistic for each sample, and plot the frequency distribution for that statistic, we would obtain an approximation of the sampling distribution.
		3. Every statistic, for example, a proportion, a mean, or a variance has a sampling distribution.
		4. Because it includes all possible sample values, the sampling distribution enables us to compare our sample result with other sample values and determine the likelihood associated with that result.
6. The Sampling Distribution of the Mean
	1. Sampling distributions are theoretical distributions, which means that they are never really observed.
	2. Constructing an actual sampling distribution would involve taking all possible random samples of a fixed size from the population.
	3. This process would be very tedious because it would involve a very large number of samples.
	4. However, to help grasp the concept of the sampling distribution, let’s illustrate how one could be generated from a limited number of samples.
	5. An Illustration
		1. The **sampling distribution of the mean** is a theoretical distribution of sample means that would be obtained by drawing from the population all possible samples of the same size.
		2. Our population is made up of 20 individuals and their incomes.
		3. From that population (Table 6.3), we now randomly draw 50 possible samples of size 3 (*N* = 3), computing the mean income for each sample and replacing it before drawing another.
		4. We repeat this process 48 more times, each time computing the sample mean and restoring the sample to the original list.
		5. This distribution is an example of a sampling distribution of the mean.
		6. Note that in its structure, the sampling distribution resembles a frequency distribution of raw scores, except that here each score is a sample mean, and the corresponding frequencies are the number of samples with that particular mean value.
	6. Review
		1. The three distinct types of distribution:
			1. The Population: We began with the population distribution of 20 individuals. This distribution actually exists. It is an empirical distribution that is usually unknown to us. We are interested in estimating the mean income for this population.
			2. The Sample: We drew a sample from that population. The sample distribution is an empirical distribution that is known to us and is used to help us estimate the mean of the population. We selected 50 samples of *N* = 3 and calculated the mean income. We generally use the sample mean (), as an estimate of the population mean (μ).
			3. The Sampling Distribution of the Mean:
				1. For illustration, we generated an approximation of the sampling distribution of the mean, consisting of 50 samples of *N* = 3.
				2. The sampling distribution of the mean does not really exist. It is a theoretical distribution.
	7. The Mean of the Sampling Distribution
		1. The sampling distribution can be described in terms of its mean and standard deviation.
		2. We use the symbol $μ\_{\overbar{Y}}$ to represent the mean of the sampling distribution. The subscript indicates the specific variable of this sampling distribution.
		3. To obtain the mean of the sampling distribution
			1. Add all the individual sample means .
			2. Divide by the number of samples (M).
		4. Thus, the mean of the sampling distribution of the mean is actually the mean of means: 
	8. The Standard Error of the Mean
		1. The standard deviation of the sampling distribution is also called the **standard error of the mean**.
		2. The standard error of the mean describes
7. How much dispersion there is in the sampling distribution
8. How much variability there is in the value of the mean from sample to sample: .
	* 1. This formula tells us that the standard error of the mean is equal to the standard deviation of the population *σ* divided by the square root of the sample size (*N*).
9. The Central Limit Theorem
	1. Central limit theorem: It states that if all possible random samples of size *N* are drawn from a population with a mean μ and a standard deviation σ, then as *N* becomes larger, the sampling distribution of sample means becomes approximately normal, with mean  equal to the population mean and a standard deviation equal to 
	2. When the population distribution is skewed, we can still assume that the sampling distribution of the mean is normal, given random samples of large enough size.
	3. The central limit theorem also assures us that (a) as the sample size gets larger, the mean of the sampling distribution becomes equal to the population mean and (b) as the sample size gets larger, the standard error of the mean (the standard deviation of the sampling distribution of the mean) decreases in size.
	4. The standard error of the mean tells how much variability in the sample estimates there is from sample to sample.
		1. The smaller the standard error of the mean, the closer (on average) the sample means will be to the population mean.
		2. Thus, the larger the sample, the more closely the sample statistic clusters around the population parameter.
	5. The Size of the Sample
		1. A general rule of thumb is that when *N* is 50 or more, the sampling distribution of the mean will be approximately normal regardless of the shape of the distribution.
		2. However, we can assume that the sampling distribution will be normal even with samples as small as 30 if we know that the population distribution approximates normality.
	6. The Significance of Sampling Distribution and Central Limit Theorem
		1. When we selected different samples, we found each time that the sample mean differed from the population mean.
			1. These discrepancies are due to sampling errors.
		2. The solution to the dilemma lies in the sampling distribution and its properties.
10. Because the sampling distribution is a theoretical distribution that includes all possible sample outcomes, we can compare our sample outcome with it and estimate the likelihood of its occurrence.
	* 1. Our knowledge is based on what the central limit theorem tells us about the properties of the sampling distribution of the mean.
		2. If our sample size is large enough (at least 50 cases), most sample means will be quite close to the true population mean.
		3. It is highly unlikely that our sample mean would deviate much from the actual population mean.
		4. In all normal curves, a constant proportion of the area under the curve lies between the mean and any given distance from the mean when measured in standard deviation units or *Z* scores.
11. We can find this proportion in the standard normal table (Appendix B).
	* 1. Knowing that the sampling distribution of the means is approximately normal, with a mean $μ\_{\overbar{Y}}$ and a standard deviation  (the standard error of the mean), we can use Appendix B to determine the probability that a sample mean will fall within a certain distance measured in standard deviation units or *Z* scores (of $μ\_{\overbar{Y}}$ or μ.)
12. Statistics in Practice: The 2016 U.S. Presidential Election
	1. There are numerous applications of the central limit theorem in research, business, medicine, and popular media. The data are all derived from relatively small random samples taken from considerably larger and varied populations.
		1. The data have consequences, informing our understanding of the social world, influencing decisions, shaping social policy, and predicting social behavior.
	2. Election of President Trump in 2016 was not through the popular vote, and the election outcome shocked many who had conducted preelection polls.
	3. Multiple reasons for the inaccuracy of the polls:
13. Nonresponse bias: occurs when certain kinds of people systematically do not respond to surveys despite equal opportunity outreach.
14. Lack of honesty by survey respondents: respondents worried about socially undesirable response of saying they would vote for Trump.
15. Pollsters inaccurately predicting which likely voters would actually vote.
	1. Research methodologists will study what happened in the 2016 election to improve future election predictions.