STATISTICS IN EVERYDAY LIFE

The GSS has been administered since 1972 and this permits us to track change in our society over time. As one illustration, data from the GSS demonstrate a considerable softening in the tendency of Americans to condemn homosexuality. In the 1970s, about 72% of respondents felt that homosexuality was "always wrong." In the most recent administration of the GSS in 2010, that percentage had fallen to about 45%.

- The mean of the sampling distribution will be the same value as the mean of the population. If *all* adult Americans have completed an average of 13.5 years of schooling ($\mu = 13.5$), the mean of the sampling distribution will also be 13.5.
- The standard deviation (or standard error) of the sampling distribution is equal to the population standard deviation (σ) divided by the square root of N.

Thus, the theorems tell us the statistical characteristics of this distribution (shape, central tendency, and dispersion), and this information allows us to link the sample to the population.

How does the sampling distribution link the sample to the population? The fact that the sampling distribution will be normal when N is large is crucial. This means that more than two-thirds (68%) of all samples will be within ± 1 Z score of the mean of the sampling distribution (which is the same value as the mean of the population), about 95% are within ± 2 Z scores, and so forth. We do not (and cannot) know the actual value of the mean of the sampling distribution, but we do know that the probabilities are very high that our sample statistic is approximately equal to this parameter. Similarly, the theorems give us crucial information about the mean and standard error of the sampling distribution that we can use—as you will see in the remainder of this chapter and the other chapters in this part—to link information from the sample to the population.

To summarize, our goal is to infer information about the population (in the case of the GSS, all adult Americans). When populations are too large to test, we use information from randomly selected samples—carefully drawn from the population of interest—to estimate the characteristics of the population. In the case of the GSS, the full sample for 2010 consists of about 2,000 adult Americans who have responded to the questions on the survey. The sampling distribution—the theoretical distribution whose characteristics are defined by the theorems—links the known sample to the unknown population.

Symbols and Terminology

In the following chapters, we will be working with three entirely different distributions. Furthermore, we will be concerned with several different kinds of sampling distributions—including the sampling distribution of sample means and the sampling distribution of sample proportions.

TABLE 6.1 Symbols for Means and Standard Deviations of Three Distributions

	Mean	Standard Deviation	Proportion
1. Samples	$\overline{X}$	s	P_s
2. Populations	μ	σ	P_u
3. Sampling distributions			
Of means	$\mu_{\overline{X}}$	$\sigma_{\overline{X}}$	
Of proportions	μ_p	σ_p	

To distinguish among these various distributions, we will often use symbols; for quick reference. Table 6.1 presents the symbols that will be used for the sampling distribution. Basically, the sampling distribution is denoted with Greek letters that are subscripted according to the sample statistic of interest.

Note that the mean and standard deviation of a sample are denoted with English letters ($\overline{X}$ and s), while the mean and standard deviation of a population are denoted with the Greek-letter equivalents (μ and σ). Proportions calculated on samples are symbolized as P-sub-s (s for sample), while population proportions are denoted as P-sub-u (u for "universe" or population). The symbols for the sampling distribution are Greek letters with English-letter subscripts. The mean and standard deviation of a sampling distribution of sample means are $\mu_{\overline{X}}$ ("mu-sub-x-bar") and $\sigma_{\overline{X}}$ ("sigma-sub-x-bar"). The mean and standard deviation of a sampling distribution of sample proportions are μ_p ("mu-sub-p") and σ_p ("sigma-sub-p").

Introduction to Estimation

The object of this branch of inferential statistics is to estimate population values or parameters from statistics computed from samples. Although the mathematical techniques may be new to you, you are certainly familiar with their most common applications. Polls and surveys on every conceivable issue—from the sublime to the trivial—have become a staple of mass media and popular culture. The techniques you will learn here are essentially the same as those used by the most reputable, sophisticated, and scientific pollsters.

The standard procedure for estimating population values is to construct **confidence intervals**—a mathematical statement that says that the parameter lies within a certain interval or range of values. For example, a confidence interval estimate might say "68% ±3%—or between 65% and 71%—of Americans approve of the death penalty."[3] In the media, the central value of the interval (68% in this case) is usually stressed, but it is important to realize that the population parameter (the percentage of *all* Americans who support capital punishment) could be anywhere between 65% and 71%.

[3]This estimate is based on the General Social Survey GSS, 2010, which was administered to a representative sample of adult residents of the United States.

Estimation Selection Criteria

Estimation procedures are based on sample statistics. Which of the many available sample statistics should be used? Estimators can be selected according to two criteria: **bias** and **efficiency**. Estimates should be based on sample statistics that are unbiased and relatively efficient. We cover each of these criteria separately.

Bias. An estimator is unbiased *if the mean of its sampling distribution is equal to the population value of interest.* We know from the theorems presented earlier in this chapter that sample means conform to this criterion. The mean of the sampling distribution of sample means (which we will note symbolically as $\mu_{\overline{X}}$) is the same as the population mean (μ).

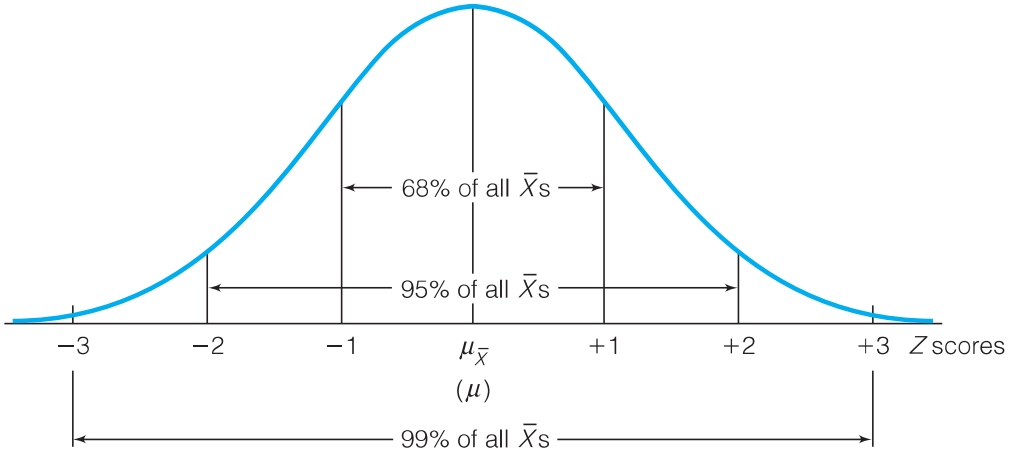
Sample proportions (P_s) are also unbiased. That is, if we calculate sample proportions from repeated random samples of size N and then array them in a line chart, the sampling distribution of sample proportions will have a mean (μ_p) equal to the population proportion (P_u). Thus, if we are concerned with coin flips and sample honest coins 10 at a time ($N = 10$), the sampling distribution will have a mean equal to 0.5, which is the probability that an honest coin will be heads (or tails) when flipped. All statistics other than sample means and sample proportions are biased (that is, have sampling distributions with means not equal to the population value).[4]

Because sample means and proportions are unbiased, we can determine the probability that they lie within a given distance of the population values we are trying to estimate. To illustrate, consider a specific problem. Assume that we wish to estimate the average income of a community. A random sample of 500 households is taken ($N = 500$) and a sample mean of $45,000 is computed. In this example, the population mean is the average income of *all* households in the community and the sample mean is the average income for the 500 households that happened to be selected for our sample. Note that we do not know the value of the population mean (μ)—if we did, we would not need the sample—but it is μ that we are interested in. The sample mean of $45,000 is important primarily insofar as it can give us information about the population mean.

The two theorems presented earlier in this chapter give us a great deal of information about the sampling distribution of all possible sample means in this situation. Because N is large ($N > 100$), we know that the sampling distribution is normal and that its mean is equal to the population mean. We also know that all normal curves contain about 68% of the cases (the cases here are sample means) within ± 1 Z, 95% of the cases within ± 2 Zs, and more than 99% of the cases within ± 3 Zs of the mean. Remember that we are discussing the sampling distribution here—the distribution of all possible sample outcomes or, in this instance, sample means. Thus, the probabilities are very good (approximately 68 out of 100 chances) that our sample mean of $45,000 is within ± 1 Z, excellent (95 out of 100) that it is within ± 2 Zs, and overwhelming (99 out of 100) that it is within ± 3 Zs

[4]In particular, the sample standard deviation (s) is a biased estimator of the population standard deviation (σ). As you might expect, there is less dispersion in a sample than in a population, and as a consequence, s will underestimate σ. However, as we shall see, sample standard deviation can be corrected for this bias and still serve as an estimate of the population standard deviation for large samples.

FIGURE 6.4 Areas Under the Sampling Distribution of Sample Means



of the mean of the sampling distribution (which is the same value as the population mean). These relationships are graphically depicted in Figure 6.4.

If an estimator is unbiased, it is almost certainly an accurate estimate of the population parameter (μ in this case). However, in less than 1% of the cases, a sample mean will be more than ± 3 *Z*s away from the mean of the sampling distribution (very inaccurate) by random chance alone. We literally have no idea if our particular sample mean of $45,000 is in this small minority. However, we do know that the odds are high that our sample mean is considerably closer than ± 3 *Z*s to the mean of the sampling distribution and, thus, to the population mean.

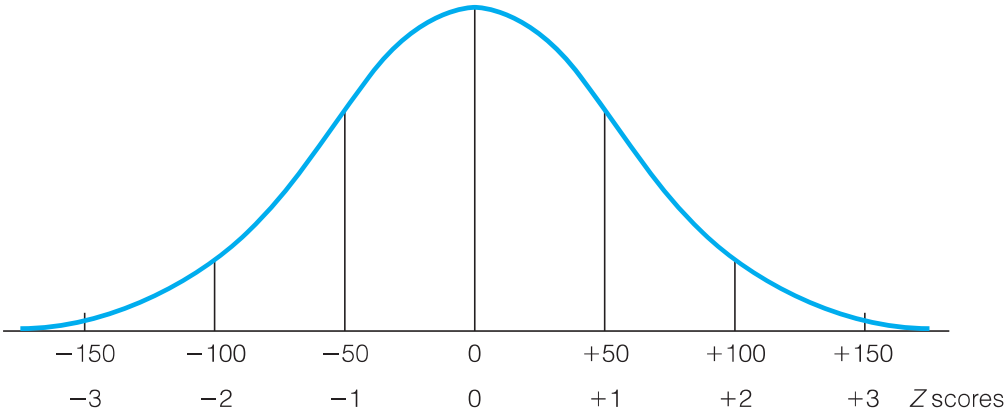
Efficiency. The second desirable characteristic of an estimator is efficiency, which is the extent to which the sampling distribution is clustered about its mean. Efficiency or clustering is essentially a matter of dispersion—the topic of Chapter 4 (see Figure 4.1). The smaller the standard deviation of a sampling distribution, the greater the clustering and the higher the efficiency. Remember that the standard deviation of the sampling distribution of sample means—or the standard error of the mean—is equal to the population standard deviation divided by the square root of *N*. Therefore, the standard deviation of the sampling distribution is an inverse function of *N* ($\sigma_{\bar{X}} = \sigma/\sqrt{N}$). As the sample size increases, $\sigma_{\bar{X}}$ will decrease. We can improve the efficiency (or decrease the standard deviation of the sampling distribution) for any estimator by increasing the sample size.

An example should make this clearer. Consider two samples of different sizes:

Sample 1	Sample 2
$\bar{X}$ = $45,000	$\bar{X}$ = $45,000
N_1 = 100	N_2 = 1,000

Both sample means are unbiased, but which is the more efficient estimator? Consider Sample 1 and assume for the sake of illustration that the population

FIGURE 6.5 A Sampling Distribution with $N = 100$ and $\sigma_{\bar{X}} = \$50.00$

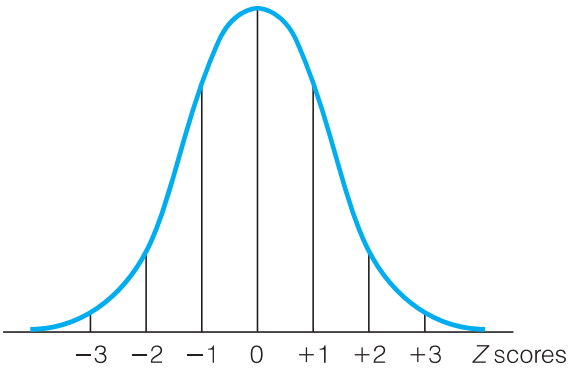


standard deviation (σ) is \$500.[5] In this case, the standard deviation of the sampling distribution of all possible sample means with an N of 100 would be $\sigma/\sqrt{N}$ or $500/\sqrt{100}$ or \$50.00. For Sample 2, the standard deviation of all possible sample means with an N of 1,000 would be much smaller. Specifically, it would be equal to $500/\sqrt{1{,}000}$, or \$15.81.

Sampling distribution 2 is much more clustered than sampling distribution 1. In fact, distribution 2 contains 68% of all possible sample means within ± 15.81 of μ, while distribution 1 requires a much broader interval of ± 50.00 to do the same. The estimate based on a sample with 1,000 cases is much more likely to be close to the population parameter than is the estimate based on a sample of 100 cases. Figures 6.5 and 6.6 illustrate these relationships graphically.

The key point to remember is that the standard deviation of all sampling distributions is an inverse function of N: the larger the sample, the greater the clustering and the higher the efficiency. In part, these relationships between the sample size and the standard deviation of the sampling distribution do nothing more than underscore our commonsense notion that much more confidence can be placed in large samples than in small (as long as both have been randomly selected).

FIGURE 6.6 A Sampling Distribution with $N = 1{,}000$ and $\sigma_{\bar{X}} = \$15.81$



[5]In reality, of course, the value of σ would be unknown.