6.3: Sampling Distributions

In previous chapters we have focused on how to summarize data from samples by looking at one sample at a time. For example we computed means, standard deviations, and even z-scores to summarize a sample's distribution (through the mean and standard deviations) and to estimate the expected locations and probabilities of individual raw scores within a distribution (through transforming to and using z-scores). The practices connects out to broader assumptions about populations and how data from samples may represent those populations.

In statistics, samples are used to estimate populations based on an assumption of how the pattern of data from many samples can represent populations. To understand this, we must first consider a concept known as a sampling distribution and how this is connected to an idea known as the Central Limit Theorem.

Let's zoom out from raw scores as the unit of analysis to sample means. Suppose that instead of considering raw scores for many cases as a way to estimate a variable, we used samples as the cases and represented them with their means. Another way to say this is that the means from many samples would be used in place of raw scores for computations. The mean of those means and the standard deviation of those means could then be used to construct things the normal curve and to estimate z-scores and probabilities to represent the population of interest. Patterns and variation in data for a variable (which would be represented in the normal curve) are referred to as *distributions.* When we use data from one sample, we refer to the patterns and dispersion as the *distribution of sample data.* When we zoom out and use means in place of raw scores, we refer to the patterns and variation as a *sampling distribution.*

Here is the boiled down explanation of what is assumed about sampling distributions: Samples are drawn from populations and are used to estimate population means. If this were to be done with replacement (meaning the full population is being sampled from each time) and a sufficient number of random samples of the population are taken, it would be called the sampling distribution. Thus, a **sampling distribution** is like a data set but with sample means in place of individual raw scores. When these samples are drawn randomly and with replacement, most of their means are expected to be fairly close to the true population mean and few are expected to be far from the true population mean. In fact, if samples are gathered from a population over and over again, the distribution of those sample means is expected to form a normal distribution the same way individual raw scores are expected to within any given sample. In this way, the distribution of many sample means is essentially expected to recreate the actual distribution of scores in the population if the population data are normal. However, even if the data in the population are skewed or are randomly generated, the sampling distribution is expected to be normal. Thus, **Central Limit Theorem** states that the sampling distribution will tend to be normal in many situations if a sufficient number of samples of sufficient size are drawn randomly and with replacement. Notice that, once again, we are stating what tends to, or is likely, to happen but not what is guaranteed. Thus, Central Limit Theorem states what is probable or tends to be true rather than what is absolute or guaranteed to occur every time.

Standard Error

Sampling distributions have means and something known as a standard error (*SE*) which together function like the mean and standard deviation (*SD*) do for individual samples. The distinction between the means is that the sampling distribution uses a mean of sample means whereas a sample uses a mean of raw scores. Both the *SE* and *SD* describe how far things tended to deviate from a mean. Thus, each is estimating how much error there is when using its respective mean to estimate something. However, these differ in what they are trying to estimate and, thus, what the error is in regards to. Specifically, *SD* focuses on how far raw scores in a sample tended to be from their mean while *SE* is used to estimate how far the mean of a sampling distribution is from the population mean it is being used to estimate.

Standard error is calculated by adjusting a *SD* by the square root of sample size using this formula:

$$SE = \frac{s}{\sqrt{n}}$$

Let's take a look at how to calculate standard error with an example.

✔ Calculating Standard Error (SE)

Given:

s = 9.3859063 …

n = 22

Find SE

Solution

$$SE = \frac{s}{\sqrt{n}}$$
$$SE = \frac{9.3859\ldots}{\sqrt{22}}$$
$$SE = \frac{9.3859\ldots}{4.6904\ldots}$$
$$SE \approx 2.0010\ldots$$

The SE can be rounded to the hundredths place and summarized as: SE = 2.00.

Sampling Error

Sampling error refers to the amount of error (or difference) between a sample statistic and the parameter in the population the sample represents. Each time a sample is drawn and a statistic such as the mean is computed, it may be discrepant from the population mean. However, as the sample size increases, more of the population is being represented by the sample and, thus, the less sampling error is expected. Standard error is a specific calculation that is used to estimate the sampling error between any one sample mean and its population mean. Therefore, sampling error refers to the broader idea that sample statistics (such as the mean) are rarely perfect representations of population parameters while standard error estimates the amount a sample mean deviates from the population mean it is being used to represent. Because population means are rarely known, standard errors are a necessary consideration when using sample means to estimate unknown population parameters.

Though there is much more that can be said about sampling distributions, Central Limit Theorem, standard errors, and sampling error, this boiled down review focused on the attributes and scenarios most applicable to the level of understanding required to move forward to probability and hypothesis testing.

Standard errors (SE)

Standard errors (SE) summarizes the average deviation between sample means ($\bar{x}$) and the population means (μ) they are being used to estimate.

SE is calculated by dividing standard deviation (s) by the square root of the sample size (n).

The symbols $\sigma_{\bar{x}}$ or σ_M are sometimes used in place of SE to emphasize that the standard error of sampling means is an adjusted version of standard deviation.

Reading Review 6.2

1. What is a sampling distribution?
2. What does Central Limit Theorem state should occur when a sufficient number of samples of sufficient size are drawn randomly and with replacement?
3. Which statistic is calculated to estimate sampling error?