**Sampling Distributions**

One of the most important concepts in inferential statistics is that of the sampling distribution. That is because the use of sampling distributions is what allows us to make “probability” statements in inferential statistics.

* A sampling distribution is defined as “The theoretical probability distribution of the values of a statistic that results when all possible random samples of a particular size are drawn from a population.” (For simplicity you can view the idea of “all possible samples” as taking a million random samples. That is, just view it as taking a whole lot of samples!)
* One specific type of sampling distribution is called the sampling distribution of the mean. If you wanted to generate this distribution through the laborious process of doing it by hand (which you would NOT need to do in practice), you would randomly select a sample, calculate the mean, randomly select another sample, calculate the mean, and continue this process until you have calculated the means for all possible samples. This process will give you a lot of means, and you can construct a line graph to depict your sampling distribution of the mean (e.g., see Figure 19.1).
* The sampling distribution of the mean is normally distributed (as long as your sample size is about 30 or more for your sampling).
* Also, note that the mean of the sampling distribution of the mean is equal to the population mean! That tells you that repeated sampling will, over the long run, produce the correct mean. The spread or variance shows you that sample means will tend to be somewhat different from the true population mean in most particular samples.

Although I just described the sampling distribution of the mean, it is important to remember that a sampling distribution can be obtained for any statistic. For example, you could also obtain the following sampling distributions:

* Sampling distribution of the percentage (or proportion).
* Sampling distribution of the variance.
* Sampling distribution of the correlation.
* Sampling distribution of the regression coefficient.
* Sampling distribution of the difference between two means.

The standard deviation of a sampling distribution is called the standard error. In other words, the standard error is just a special kind of standard deviation and students learned what a standard deviation was in the last chapter.

* The smaller the standard error, the less the amount of variability present in a sampling distribution.

It is important to understand that researchers do not actually empirically construct sampling distributions! When conducting research, researchers typically select only one sample from the population of interest; they do not collect all possible samples.

* The computer program that a researcher uses (e.g., SPSS and SAS) uses the appropriate sampling distribution for you.
* The computer program will look at the type of statistical analysis you select (and also consider certain additional information that you have provided, such as the sample size in your study), and then the statistical program selects the appropriate sampling distribution.
* (It is kind of like the Greyhound Bus analogy: Leave the driving to us … SPSS will take care of generating the appropriate sampling distribution for you if you give it the information it needs.)

So please remember that the idea of sampling distributions (i.e., the idea of probability distributions obtained from repeated sampling) underlies our ability to make probability statements in inferential statistics.

Now, I am going to cover the two branches of inferential statistics (i.e., estimation and hypothesis testing) that were shown in Figure 19.1: estimation and hypothesis testing.

**Estimation**

The key estimation question is “Based on my random sample, what is my estimate of the population parameter?”

* The basic idea is that you are going to use your sample data to provide information about the population.

There are actually two types of estimation.

* They can be first understood through the following analogy: Let us say that you take your car to your local car dealer's service department and you ask the service manager how much it will cost to repair your car. If the manager says it will cost you $500 then she is providing a point estimate. If the manager says it will cost somewhere between $400 and $600 then she is providing an interval estimate.

In other words, a point estimate is a single number, and an interval estimate is a range of numbers.

* A point estimate is the value of your sample statistic (e.g., your sample mean or sample correlation), and it is used to estimate the population parameter (e.g., the population mean or the population correlation).
* For example, if you take a random sample from adults living in the United States and you find that the average income for the people in your sample is $45,000, then your best guess or your point estimate for the population of adults in the U.S. will be $45,000.

In the above example, you used the value of the sample mean as the estimate of the population mean.

* Again, whenever you engage in point estimation, all you need to do is to use the value of your sample statistic as your “best guess” (i.e., as your estimate) of the (unknown) population parameter.

Oftentimes, we like to put an interval around our point estimates so that we realize that the actual population value is somewhat different from our point estimate because the sampling error is always present in sampling.

* An interval estimate (also called a confidence interval) is a range of numbers inferred from the sample that has a known probability of capturing the population parameter over the long run (i.e., over repeated sampling).
* See Figure 19.2 for a picture of 20 different confidence intervals randomly jumping around the population mean from sample to sample. Here it is for your convenience:
* The “beauty” of confidence intervals is that we know their probability (over the long run) of including the true population parameter. (You cannot do this with a point estimate.)
* Specifically, if you have the computer provide you with a 95 percent confidence interval (based on your data), then you will be able to be “95% confident” that it will include the population parameter. That is, your “level of confidence” is 95%.
* For example, you might take the point estimate of annual income of U.S. adults of $45,000 (used earlier as a point estimate) and surround it by a 95% confidence interval. You might find that the confidence interval is $43,000 to $47,000. In this case, you can be “95% confident” that the average income is somewhere between $43,000 and $47,000.
* If you have the computer program give you a 99% confidence interval, then you can be “99% confident” that the confidence interval provided will include the population parameter (i.e., it will capture the true parameter 99% of the time in the long run).

You might ask: So why do not we just use 99% confidence intervals rather than 95% intervals, since you will make fewer mistakes?

* The answer is that for a given sample size, the 99% confidence interval will be wider (i.e., less precise) than a 95% confidence interval. For example, the interval $40,000 to 50,000 is wider than the interval $43,000 to $47,000.
* 95% confidence intervals are popular with many researchers. However, you may, at times, want to use other confidence intervals (e.g., 90% confidence intervals or 99% confidence intervals).

**Hypothesis Testing**

Hypothesis testing is the branch of inferential statistics that is concerned with how well the sample data support a null hypothesis and when the null hypothesis can be rejected in favor of the alternative hypothesis.

* First note that the null hypothesis is usually the prediction that there is no relationship in the population.
* The alternative hypothesis is the logical opposite of the null hypothesis and says there is a relationship in the population.
* We use hypothesis testing when we expect a relationship to be present; in other words, we usually hope to “nullify” the null hypothesis and tentatively accept the alternative hypothesis. (Note: if you expect the null to be true, you can use the estimation approach described in this chapter; several additional procedures for this special case are discussed in Shadish, Cook, and Campbell’s book *Experimental and Quasi-Experimental Designs*, 1902, pp. 52–53)
* Here is the key question that is answered in hypothesis testing: “*Is the value of my sample statistic unlikely enough (assuming that the null hypothesis is true) for me to reject the null hypothesis and tentatively accept the alternative hypothesis*?”
* Note that it is the null hypothesis that is directly tested in hypothesis testing (not the alternative hypothesis).