Statistics Background

Statistics is a mathematical field that allows us to summarize and analyze data. The two major subdivisions are **descriptive statistics** and **inferential statistics**. In descriptive statistics, the main goal is data reduction. That is, large amounts of data are summarized, typically with graphs or tables and by calculating measures of central tendency and variability. Although these summaries are convenient for simplifying data and making patterns more obvious, it should be recognized that much of the detail is lost during summarizing. Descriptive statistics are employed when the whole population can be sampled.

When the whole population cannot be sampled (the typical situation), a sample (or several) is withdrawn from the population and inferences about the population are made from the characteristics of the sample. Thus, it is assumed that a sample is a good representative of the population. However, this is not always the case and statistics also allows us to decrease our uncertainty about drawing conclusions from incomplete data.

Measures of Central Tendency

A **distribution** is created when data variables are arranged in order from highest to lowest, typically represented graphically as a **frequency distribution**, plotting the value of the variable versus the frequency of that value’s occurrence. It was shown quite a long time ago that if the number of data points is very large, the distribution usually becomes a normal or Gaussian distribution (the bell-shaped curve). Researchers can learn a lot from these plots, including the shape of the distribution, the range of values, and the most common value. Measures of **central tendency**, that is the values that the distribution seems to cluster around, are some of the most commonly calculated statistics.

The **mean** is one of the most used statistics in all manners of research. It is the arithmetic average of values, calculated as the sum of all values divided by the number of values. Although the mean provides a simple summary of a distribution, it doesn’t indicate anything about the range of values. The distribution may have a very tight distribution around the mean or may be so spread out that a peak is hard to identify. Another shortcoming of the mean is that it is sensitive to extreme values. A single outlying data point can skew the distribution so that the mean no longer represents the peak of the curve. Because of this, other descriptors of the central tendency may actually be more useful.

The **median** is the midpoint of the distribution, with half of the values on either side of it. Another way to say this is that the median represents the 50th percentile. The median can tell us about the shape of the distribution. In a normal distribution, the mean and median (and mode) are the same. If the distribution is skewed, the mean is closer to the mode than is the median. The median is usually the best measure of central tendency in a skewed distribution, for example the salaries in the NBA where a few people earn much, much more than the others.

The Central Limit Theorem

The **Central Limit Theorem** states that the means of independently drawn random samples will be approximately normally distributed, if the sample size is large enough. In other words, even if a
distribution is not normal, the distribution of the means will be. Means calculated from observation sets containing thirty sample values is considered large enough to produce a normal distribution of means from non-normally distributed data, but 10 sample values per observation set may be enough in many cases. The mean of a sampling distribution of the means is considered the same as the population mean and is often called the **expected value**. When we calculate a mean from random samples drawn from a population, we expect that mean to be the same as the population mean. The dispersion (standard deviation) of the sampling distribution of the means is called the **standard error** because it is drawn from a sampling distribution rather than a data distribution. It represents how confident we should be that a sample mean represents the population mean. The standard error is equivalent to the population standard deviation divided by the square root of the number of samples in each observation set. The standard error is smaller than the population standard deviation so distribution of the means has less dispersion than the distribution of the population. Thus, a sampling distribution of the means can provide more accurate inferences about a population than a distribution from a large pool of data.

In practice, sampling distributions of the mean are usually not constructed but knowledge of the Central Limit Theorem increases our confidence when making an inference based on a single sampling. We know that if we take several sets of samples from a population, they would have a normal distribution. So, a mean from a single set of samples taken from a population represents one mean in a normal distribution. Based on the Central Limit Theorem it is possible to calculate the probability that a sample or predicted outcome is significantly different from the population mean, but that’s a discussion for another time.

**Hypothesis Testing**

We know that a hypothesis is a testable prediction that is consistent with specific observations. With testing, it is possible to show that a hypothesis is false but not true. At best, we refuse to reject a hypothesis based on current data but recognize that as yet unexamined conditions may show the hypothesis to be false sometime in the future. In statistical hypothesis testing, it is common to reduce the question at hand to two outcomes (e.g., A and B are the same, A and B are not the same). We can state these outcomes as the **null hypothesis** ($H_0$), or the hypothesis of no difference, and the **alternative hypothesis** ($H_a$).

Two of the most common statistical tests used for determining if a significant difference exists between two or more samples is the $t$-test and ANOVA (analysis of variance). The $t$-test is a powerful and robust test for determining if two populations have different means when samples are independent and random and the measured variable is continuous and normally distributed. ANOVA is more appropriate (and less error prone) than running multiple $t$-tests when there are more than two samples to compare. In ANOVA there are two types of variance to consider: error variance (within-groups variance) and treatment variance (between-groups variance). When samples are independent and random, the within-groups variance should be the same. Therefore, any difference between the variances of the measurements would be due to the treatment. ANOVA can indicate a difference between the treatments but additional tests are necessary to find which treatments are significantly different. A level of significance must be selected for statistical hypothesis testing, designated as an alpha level. An alpha value of 0.05 is the level usually selected as it offers a good compromise to avoid Type I (rejecting a null hypothesis that is true) or Type II (not rejecting a null hypothesis that is false) errors.