Sampling Theory

Sampling theory is a branch of statistics dedicated to providing appropriate mathematical and theoretical foundations for the use of statistics to describe a set of data, or to predict outcomes with data, or make inference about a sample or its population from a sample of the population. The following are brief summaries of the essential knowledge base required to handle and use sampling data or statistics appropriately.

Population versus Sample: A population is a group of data set that is alike in one or more characteristics as *defined by the researcher*. A population may be all leadership students working on their dissertations. A sample is group of data selected from the population and is smaller than the size of the population. Samples may be selected many different ways: random, stratified, or cluster.

Statistics versus Parameters: When a statistics is computed from a sample of the population, such as the mean and standard deviation, we call such measures **statistics**. When a statistics, such as the mean and standard deviation, is computed from population (data from the entire population), we call the statistics a **parameter.** The standard deviation when computed from a sample is denoted by the symbol ***S***; while the standard deviation computed from the entire population is denoted by the symbol σ (called, sigma).

Descriptive versus Inference statistical methodologies: The main difference between descriptive and inferential statistics is the purpose for doing the measurement, not in the kind of statistical measures used.

If you which to describe a set of data, you may draw a frequency polygon; compute the central tendency by calculating its mean, median or mode; or you may compute its standard deviation to find its variability. If, however, your purpose is to make inference about a population from the sample, you will still calculate the mean and standard deviation, but your main focus is how these statistics estimate a population mean or standard deviation.

Types of Samples:

1. **Random Samples**: A random sample is one chosen in such a manner that each data point in the population has an equal change of being selected or being in the sample. There are two ways to select a sample randomly: a. *System of Random Number Generation*: Numbered the population, 1 to *N* (*N* is total in population), use a table or system of random number generation and assign a subset of elements equal to the desired sample size randomly, and select from the population those elements or data point that matches the randomized assignment, b. *Counting-Off Procedure* (called systematic sampling): Numbered the population, 1 to *N*, let *X* be *N* divided by the sample size, Select a random starting point, and select you sample as every X^{th} labeled or numbered items from the population.

2. **Stratified Random Samples**: This sampling selection technique tries to get a represented sample from a population with varying characteristics. If a population consists of 20% males and 50% Democrats, it might be necessary to divide the population into subgroups of males and Democrats and then try to get a representative sample that is representative of the population of about 20% males and 50% Democrats.

3. **Cluster Sampling:** Often is very difficult to stratify a population on a national basis, so cluster sampling is used instead of stratified sampling. If you which to develop a math standard for college math students, it would be difficult to define a sample for all college students nationally. A cluster would be any intact unit such as states, counties, or school district. We may elect to use counties as our cluster. From each county with colleges we would gather from these college students any variables affecting our measures (e.g. Gender, age, GPA) and include these in our cluster. We would select a random sample of clusters in a first stage sampling effort, and additional random sampling would be used to select particular schools, classrooms, or even individuals within the classroom to achieve a represented sample of the population.

Sample Size: A sample is more likely to accurately describe the population characteristics is it is large. A random sample is often preferred in sampling.

Biased Samples: Any sample containing a systematic error is said to be a *biased* sample. If you wanted to measure the public views on the consumption of alcoholic beverages, your sample may be biased it you only select the opinions of persons coming out of a bar.

Incidental Sampling: This is often a sample of convenience and often the results cannot be generalized to the population. If you collect data based on a sample of college freshmen, you may not be able to use the results to generalize about college seniors.

Sampling with Replacement: Sampling that requires repeated selection from a population of which the sample is returned is called sample with replacement. Non-replacement of samples from a population and replacement sampling has different outcome probabilities and a researcher must be aware of this.

Estimating the Population Mean - Every different sample from a population gives different results. Most of the energy employed in inferential statistics goes toward estimation of population parameters. Statisticians have developed the necessary techniques to estimate the population parameters with a given degree of certainty.

Sampling Distribution of the Mean: If one takes a very large sample from the population, one may be able to come very close to estimating its true parameter, such as its mean. If one takes repeated samples from a population and measure its mean, a distribution of all those means would look like a normal distribution, of which the center or "mean" of such distribution would be very close to the population mean, the standard deviation of such mean would be the standard deviation divided by the square root of the sample size, N:

$$\frac{\sigma}{\sqrt{N}}$$

This is called the **Central Limit Theorem**: The means of repeated samples of the population is normally distributed around the "true" mean or the population with a standard deviation as above. The standard deviation of this sampling distribution of means is called the ***standard error of the mean***. We use the standard deviation of our sample to estimate this standard deviation.

Since the sampling distribution of means is normally distributed, we can use the properties of the normal curve discussed earlier to make inference of our sample mean relative to normal curve. We can say, for example, that 5% of all sample means would fall outside the same interval. -1.96 and 1.96 standard deviations from the estimate of the population mean or the probability that a sample mean falls outside the same interval (population mean +/- 1.96 standard deviations) is 0.05.

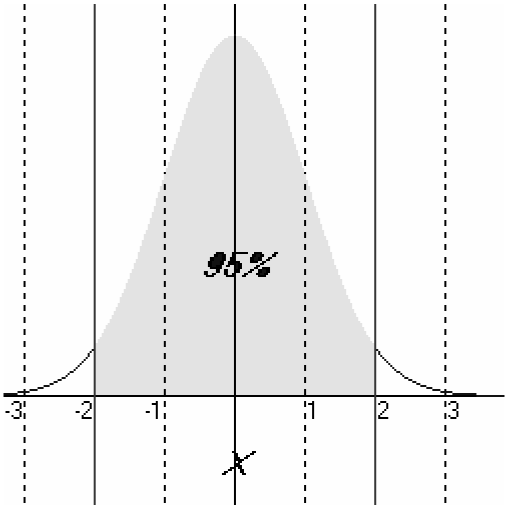

Figure 1. Probability of normal curve within z = -2 and +2.

Hypothesis Testing: If we wish to compare two means, we can ask the questions, are these two means different? One way of do this is to subtract the value of one mean from the other and see if the difference is “0”. Therefore, we can test the notion or theory that mean one equals mean two or mean one - mean two = 0?, i.e. We assume that there is no difference in the means. *This statement is called the null hypothesis.*

Null Hypothesis Assumes, H_0: $\mu_1 - \mu_2 = 0$ or $\mu_1 = \mu_2$

where $\mu_1 =$ mean 1 and $\mu_2 =$ mean 2

The hypothesis that the *null hypothesis* is not true is called the *alternate hypothesis*.

$$H_a : \mu_1 \neq \mu_2$$

Significance Level: When we test the null hypothesis, we may get a value close to zero for the difference of the means; must how small is small? How do we tell is the difference is large or small? The difference may be just due to sampling or random error? The level of significance allows statisticians and researchers to gage whether the error is due to sampling error or some assignable causes. The ***significance level*** is the probability that a result is due to sampling error, and *if this probability is small enough, we reject the notion that sampling error is the cause*. We then conclude that there is a real difference between our result and what would logically be expected by change. Traditionally these significant levels have been set at 0.05 and 0.01. The significance level is usually denoted by the *p*-value:

Reject Null Hypotheses if p =< 0.05 (5% chance)

(where, *p*-value is significance level)

Examples of Hypothesizing: We measured the height of 100 men 18 years and old, and calculated a mean height of 69.2 inches and a standard deviation of 3.0 inches. We don not really know the mean height of all men in our population but we could quest or form an opinion about what we thing the true average height. We can argue that if we measured a large number of men our average would approach that or the population. We know based on the Central Limit Theorem, that the true average height of all men is normally distributed with standard deviation of the standard error.

Test 1: let is test if the sample mean height of 69.2 is close to the hypothesized true height of 69. So we formulate a null hypothesis:

$$H_0 : M = 69$$

We can calculate the standard error as:

$$\sigma_M = \frac{S}{\sqrt{N}} = \frac{3}{\sqrt{100}} = 0.3$$

Using the z-score for a normal distribution, we calculate z:

$$z = \frac{M - 69}{\sigma_M} = \frac{69.2 - 69}{0.3} = 0.67$$

From our z-score table of probability we see that the probability of the scores or data falling outside -0.67 (0.2514) and +0.67 (0.2514 = 1 - 0.7486) is ***p*** = **0.5028**. Since $p > 0.05$ we have no reason to reject the null hypothesis that 69.2 are equal to the true mean of 69. This is a complicated way of saying, we that we would conclude that *it is entirely possible that the population mean is in fact 69* and our sample mean of 69.2 represents sampling error due to chance. This is also a round about way of saying (statisticians never said this out right) that we accept the null hypothesis.

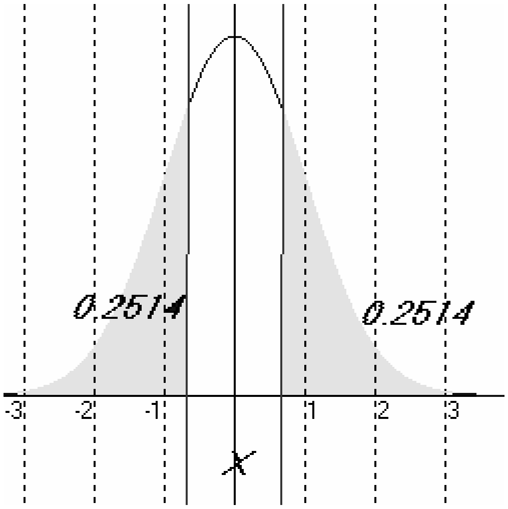

Figure 2. Probability outside normal curve for z=-0.67 and 0.67.

Test 2: let is test if the sample mean height of 69.2 is close to a hypothesized true height of 68.6. So we formulate a null hypothesis:

$$H_0 : M = 68.6$$

We can calculate the standard error as:

$$\sigma_M = \frac{S}{\sqrt{N}} = \frac{3}{\sqrt{100}} = 0.3$$

Using the z-score for a normal distribution, we calculate z:

$$z = \frac{M - 68.6}{\sigma_M} = \frac{69.2 - 68.6}{0.3} = 2$$

From our z-score table of probability we see that the probability of the scores or data falling outside -2 (0.0228) and +2 (0.0228 = 1 - 0.9772) is ***p*** = **0.0456**. Since ***p*** < 0.05 we **reject the null hypothesis** that 69.2 is equal to the true mean of 68.6.

1. Our hypothesis is incorrect that the population mean is 68.6, or

2. The population mean could really be 68.6 inches and our sample mean of 69.2 is a rare occurrence, one of the few that we expect would deviate this much by sampling error.

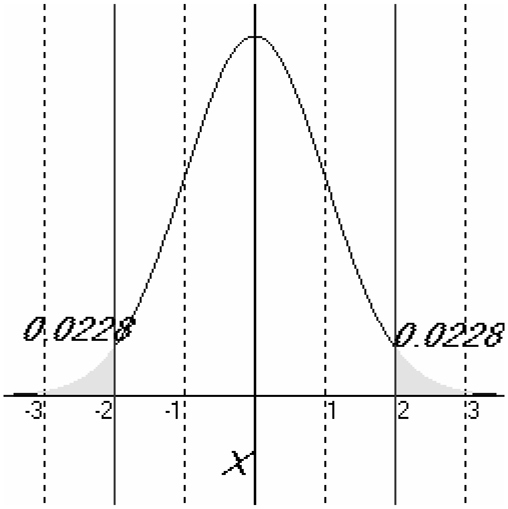

Figure 3. Probability outside normal curve when z = -2 and 2.

Notice the two ways of looking at the result when we reject the null hypothesis.

1. The result favors the null hypothesis or we do not reject the null hypothesis
2. We reject the null hypothesis and conclude that the alterative hypothesis is true