**Module 2 Review**

**Statistical Distribution** - The distribution of a statistical data set (or a population) is a listing or function showing all the possible values (or intervals) of the data and how often they occur. When a distribution of categorical data is organized, you see the number or percentage of individuals in each group.

**Probability Mass Function(PMF)** - PMF is used when the solution that you need to come up with would range within numbers of discrete random variables

**Probability Density Function(PDF)** - PDF is used when you need to come up with a range of continuous random variables.

**Cumulative Distribution Function(CDF)** - defined as the fraction of the total number of possible outcomes X (a random variable), which are less than a specific value x (a number)

**Bernoulli Distribution** - a [discrete distribution](http://mathworld.wolfram.com/DiscreteDistribution.html) having two possible outcomes labelled by n=0 and n=1 in which n=1 ("success") occurs with probability p and n=0 ("failure") occurs with probability q=1-p, where 0<p<1. It therefore has [probability density function](http://mathworld.wolfram.com/ProbabilityDensityFunction.html).

**Uniform Distribution** - each outcome is equally likely. Ex. a dice roll

**Normal Distribution** - is the classic "bell curve" with 68% of the probability mass within 1 SD of the mean, 95% within 2 SDs and 99.7% within 3 SDs

**Poisson Distribution** - can be used to display the likelihood of a given number of successes over a given time period - e.g. "how likely is it that 25 people walk into a store in a given hour?"

**Exponential Distribution** - can be used to describe the probability distribution of the amount of time it may take before a given event to occur

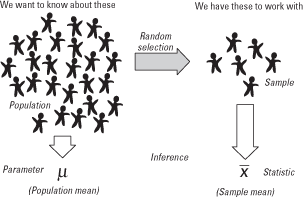
**Central Limit Theorem** - independent random variables summed together will converge to a normal distribution as the number of variables increase

**Z-Score** - the number of standard deviations you'd have to go from the mean of the normal distribution to capture the proportion of the data associated with the desired confidence level

**P-Value** - The probability of observing a test statistic at least as large as the one observed, by random chance, assuming that the null hypothesis is true. If you calculate a p-value and it comes out to 0.03, you can interpret this as saying, there is a 3% chance of obtaining the results I'm seeing when the null hypothesis is true. Simply put - the p-value is the probability that our conclusions are due to random chance. IN ADDITION - you could also use the t-statistic(computed from sample data) instead of a p-value.

**α (alpha value)** - The marginal threshold at which you're okay with rejecting the null hypothesis. If you set an alpha value of α=0.05, you're essentially saying "I'm okay with accepting my alternative hypothesis as true if there is less than a 5% chance that the results that I'm seeing are actually due to randomness." If p<α, then you reject the null hypothesis.

**Hypothesis Test** - A Hypothesis Test evaluates two mutually exclusive statements about a population to determine which statement is best supported by the sample data



**Null Hypothesis** - There is no relationship between A and B

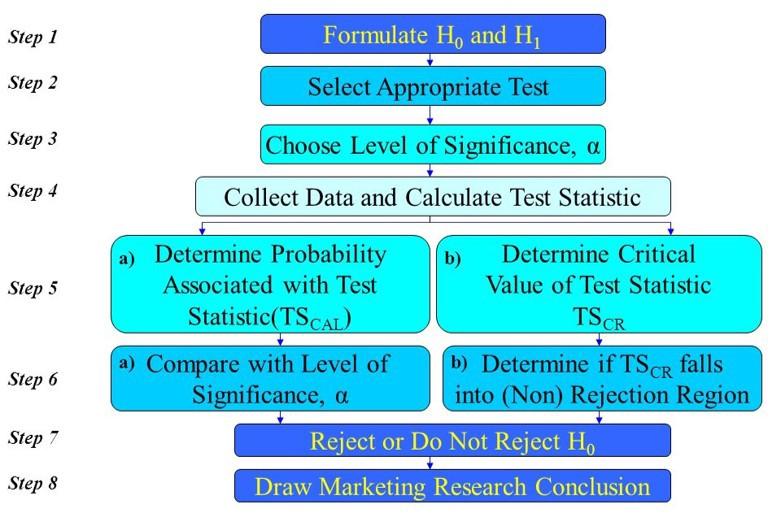
**Alternate Hypothesis** - There is a relationship between A and B(what your trying to prove)

**Confidence Interval** - a range of values so defined that there is a specified probability that the value of a parameter lies within it.

**T- Distribution** - In probability and statistics, Student's t-distribution is any member of a family of continuous probability distributions that arises when estimating the mean of a normally distributed population in situations where the sample size is small and the population standard deviation is unknown.

**T-test-Assumptions** -

1. The independent variable is categorical
2. The dependent variable is continuous
3. The dependent variable should not contain any significant outliers
4. The variances between the 2 groups are equal



**Skewness** - the degree of distortion from the symmetrical, normal distribution. A positive skew is when the tail of the right side of the distribution is longer. The mean and median are bigger than the mode in this case. The best way to determine if a distribution is skewed is visually.

**Kurtosis** - the degree of peakedness

**One-Tailed T-Test** - is when you want to know if a parameter from the treatment group is greater than (or less than) a corresponding parameter from the control group.

**Two-tailed T-test** - In statistics, a two-tailed test is a method in which the critical area of a distribution is two-sided and tests whether a sample is greater than or less than a certain range of values.

* It is used in null-hypothesis testing and testing for statistical significance.
* If the sample being tested falls into either of the critical areas, the alternative hypothesis is accepted instead of the null hypothesis.
* By convention two-tailed tests are used to determine significance at the 5% level, meaning each side of the distribution is cut at 2.5%.

**ANOVA -** An ANOVA test is a way to find out if a survey or experiment results are [significant](https://www.statisticshowto.datasciencecentral.com/what-is-statistical-significance/) by comparing the variances between groups

**F-statistic** - test statistic for Anova, defined only for positive values typically used with positively skewed data

**Type 1 Error** - Rejecting the null hypothesis when it is actually true is called a type 1 error. Committing a type 1 error is a false positive because we end up recommending something that does not work.

**Type 2 Error** - A type 2 error occurs when we do not reject the null hypothesis when it is actually false. This is a false negative because we end up doing nothing when we should have taken action.

**Effect size** - a quantitative measure of the magnitude of the experiments effect. The larger the effect size the stronger the relationship between two variables. The correlation between 2 variables; the regression coefficient; the difference in means are all examples of effect size.

**Cohen’s d effect size:** Cohen’s d is known as the difference of two population means and it is divided by the standard deviation from the data. It is used to measure the effect size of 2 samples by comparing the means and SD. A cohen’s d value > 0.5 is considered to have a medium effect and a value > .8 has a strong effect

