**Stats Vocabulary**

* **Standard Deviation** – average squared distance from the mean
  + % of data that usually falls within number of deviations is as follows:   
    ~ 68% between -1 and 1  
    ~ 95% between -2 and 2  
    ~ 99.7% between -3 and 3
* **Standardized test score** – ex z score: measures distance of a certain point to the mean in terms of standard deviation units
* **Confidence Interval** – creating an interval (range) in which we are x% confident that the true population test statistic lies
* **Statistical efficiency** – the degree of the variability of an estimate relative to other estimates of the same quantity
* **Sampling distribution** – the distribution of the statistic of interest when that statistic is calculated from random samples of the data under the same set of conditions
* **Standard Error** – sample to sample variation (if hypothetically a study were repeated multiple times then compared) using standard deviation of sample means
* **Sampling Error** – the random difference between an estimate and the parameter estimates (sample vs population)
* **Sample Variance** – describes the mean of the squared deviations between the scores and the mean of scores. Expresses variability (square root of this is standard deviation)
* **Margin of error** – the points at either end of the confidence interval ex if 95% of sample scores fall between 47% and 53% test result (if sampling voter party in this case) the margin of error would be 6% (53% - 47%)
* **Covariance** – measures how changes in one variable can predict changes in another aka strength of linear relationship
  + tests HOW variables interact AKA the level of association between 2 variables (not normalized)
  + positive result means variables move same direction
  + negative result means they move opposite to each other
* **Correlation** – measure of the strength of the linear association or relation between 2 variables, standardized measure meaning it does not depend on units
  + Ranges between -1 and 1 with symbol “r”
  + Measure of how well our linear model line fits is also between -1 and 1 symbol “R^2”
* **Test statistic** – our calculated # that we get when applying our specified test for hypothesis testing (either for example t or z)
* **Critical value** – region of rejection expressed in terms of distribution score ex z/t score of +/- 1.96
* **P-value** the probability, given the null hypothesis is true, of obtaining results as or more extreme than that which were observed
* **Type I error** – false positive ex. test positive for cancer when you don’t have it
* **Type II error** – false negative ex. test negative for cancer when you do have it

**Popular Test Uses**

* **T-test** – compares the means of 2 samples for significance
  + samples DO NOT need to be from the same population
  + use if n < 100 and population variance or standard deviation is unknown
  + **2 sample t-test (unpaired)** –Two samples from two different populations (Student’s and Welch’s
  + **Paired t-test** – Two samples from the same population, usually a “before and after” test
  + **One Sample** – compares the sample mean to a known mean
  + Any test can either have one or 2 tails
* **Z-test** – compares larger sample mean to population mean and the sample is assumed to come from the population
  + use if n > 100 and population variance or standard deviation is known
* **Chi Square** – Goodness of Fit test, used to compare categorical data
  + Observed vs Expected (usually categorical data)
* **Chi Square** – test for independence, used to determine if 2 sets of categorical variables are associated or independent
  + Ho = no association, Ha = there is an association
* **Chi Square** – test for single variance, determines if true population variance is different than hypothesized population variance by using a sample from the population
  + Ex. comparing long time measured population to current population after an adjustment was make by taking a sample from new population
* **F-test** – for 2 variances, tests difference in variances between 2 independent populations
  + Ex. NYSE variance vs. NASDAQ variance
* **ANOVA** – compares means (using variance) of 2 or more categorical samples
  + used to determine if the variation between reported outputs is the result of a particular factor or randomness
  + two-way ANOVA would be examining multiple factors of each sample such as age, groups, names
  + Ex. mean sales of grocery store vs. drug store vs. convenience store
* **Regression Analysis** – helps us investigate the relationship between 2 variables such as education and lifetime earnings
  + Expressed graphically with a scatterplot and best fit line
  + Finds the formula that will create the best line to fit the data to make future predictions
* **Linear Regression** – finds line which best describes relationship between 2 variables
  + Also used to predict a dependent variable using an independent variable
  + X axis = independent variable
  + Y axis = dependent variable
* **R^2 –** coefficient of determination, score between 0 and 1 that tells us if our regression line is a good or bad fit for our plotted data

**r** – correlation coefficient, score between -1 and 1 that shows how one variable affects the other, r = square root of R^2