**Final Project Notes**

**Wednesday - Data Simulation**

* Types of data
  + Numeric/quantitative
    - Discrete
    - Continuous
  + Categorical/qualitative
    - Nominal
    - Ordinal
  + Character
  + Factor
  + Logical
* Rnorm() plots random continuous numbers from a normal distribution
* Runif() plots positive, continuous numbers within a range
* Rpois() plots random discrete integers from a poisson distribution
* Rep() and letter[] plot character variables repeatedly, as defined by a range
* Replicate() repeats the creation of a data set simulation n times

**Thursday – t-tests, ANOVA, factor analysis, structural equation modeling**

t-test

* Compare means of two samples relative to their CI
  + Unpaired= two samples from a different population
  + Paired= two samples from the same mean
* Compare means of one sample and the population (expected) relative to their CI
* Can run a one-tale or two-tale test
* Checks if the signal (difference of means) is significant or just noise
* More variance = more noise and a lower t-score
* Larger n = more signal
* Used for hypothesis testing
  + Null= no difference in means
  + Alt= difference in means
  + Reject null if your t-value is above your critical value
  + Fail to reject null if your t-value is below your critical value
    - Critical value is calculated from df and CI using a t-value table
* Assumptions
  + Normal distribution
  + Similar variance
  + N’s should be the same (or at least similar)
  + Sample size between 20-30ish
    - If higher, go for a z-test

ANOVA

* Analysis of Variance
* Involves an f-ratio (SS= sum of squares)
* Allows you to compare more than 2 samples
* You can test if all of your given means come from the same population
  + Let’s say you have three means
    - You can test if one is so far from the other two that it is likely from a separate population
    - Or you can test if all three are so far from each other that they are all likely from different populations
  + You take all points from all three distributions and place them in one large distribution
    - You then ask: where is each mean relative to the overall distribution (x-bar 1, 2, and 3 relative to the mean of the distribution containing all three)
    - If one of your x-bar for one group is outside of the CI of your overall distribution, you can reject the null that it is from the same population as the other two means
* The null hypothesis in ANOVA is “these means come from the same population”
  + This is the variability between sample means
* You could test all sample means using pairwise t-tests (essentially the same as ANOVA)
  + Each independent test would look like this:
  + Note that the error of your CI (your alpha) will **compound** with each comparison
    - Comparing each pair means 0.95\*0.95\*0.95=0.857
      * Your CI is now 85.7% instead of 95%
      * Your alpha is now 0.143 instead of 0.05
        + This means that the spread of your distributions widens, and extreme results become more likely

This is called the variability within

* + - * + This is why you don’t run pairwise t-tests and why we have ANOVA
* ANOVA is a variability ratio (ANOVA = variability between the mean / variability within the distributions
* Variability between and within are the components of total variance (variability between + variability within = total variance)
  + If the between variance is relatively large compared to the within variance, your ratio will be larger than 1
    - If your ratio is larger than 1, your samples likely did not all come from the same population and you can reject the null
      * If , then you probably can reject the null
      * If , then you probably fail to reject the null
      * If , then you probably fail to reject the null
        + This tends to indicate that the means are very close together and or the distributions “melt” together (they have large spread and a lot of overlap)
      * In the case of k= 3, this could mean that 1 of your groups is far from the overall distribution mean, or all 3 are
        + A close up of a map

          Description automatically generatedThere is no way for you to know just from an ANOVA test
* Assumptions
  + Data errors are normally distributed
  + Equal variance between groups/treatments
    - Homogeneity of variances
    - Homoscedasticity
  + Independence of samples
    - Each sample is randomly selected and independent
      * One sample being selected does not affect the chance of another sample being selected
        + Coin tossing, for instance
  + Note that you can run a one way ANOVA (one tailed) but it is generally expected that you will run two way (two tailed) ANOVA’s