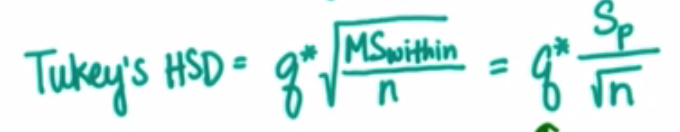
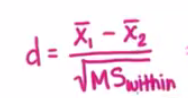
***Udacity Data Analyst Track***

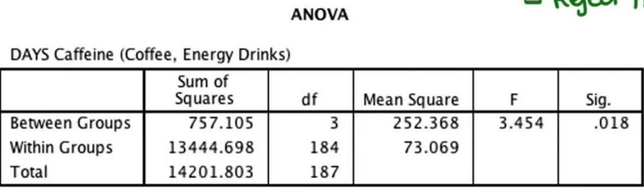
**I. Into to Inferential Stats**

9. Correlation

* Need to do additional testing to figure out which means differ from each other and why = **multiple comparisons tests**
* multiple comparisons tests are not done until 1-way ANOVA is completed
* Common test = Tukey’s Honestly Significant Difference (HSD) 🡪 evaluates significance of difference between any 2 group’s means to allow us to make pair-wise comparisons
* Calculated similarly to Margin of Error (Z \* SE, or Z \* Sigma/Sqrt(n))
* Now comparing 3 or more samples, need new statistic, the **studentized range statistic** (**q) \* Sqrt[MS(w)/n]**
* Mean square for WG variability = Pooled variance = average squared deviation of each value from its respective group’s mean
* Therefore it’s really **q = S(p) / Sqrt(n)**
* **studentized range statistic** is from a table and adjusts the whole HSD so that’s is less likely we commit a Type I Error (reject h(0) when actually true)
* As # of samples increases, q adjusts upwards (increases when more groups are being compared) = less likely to make Type I error
* Remember when sample mean was further from population mean than the margin of error, it’s unlikely to have happened by chance
* Now, if 2 samples means are further apart than the HSD, the difference is honestly significant
* 
* \* CAN ONLY COMPUTE WHEN SAMPLE SIZES ARE THE SAME
* \* n = # OF SAMPLES W/IN EACH MEAN
* Can also do Cohen’s d for multiple comparisons (measures of effect size) 🡪 d = x1 – x2 / S(p) [difference between sample means divided by pooled standard deviation for independent samples t-tests]
* Square root of Mean Square for WG variability**, Sqrt[MS(w)]**, is essentially the **S(p),** or pooled standard deviation



* Must complete 3 Cohen’s d for the 3 pairs of samples we have
* Another effect size for 1-way ANOVA is **eta squared 🡪 n^2** 🡪 tells us the proportion of total variation due to BG variability (similar to R2) = **explained variation**
* **N^2 = SS(b) / SS(ttl)** or **SS(b) / (SS(b) + SS(w))**
* Any eta squared value > 0.4 🡺 pretty large
* F is typically reported as F( dF(b), dF(w) ) = F p < alpha n^2 = .xx
* **F(2, 6) = 27 p < .05 n^2 = .90**
* Can calculate exact p-values w/ software, which we use w/ large samples
* When using software, don’t about intermediate steps, only about comparing p-value w/ alpha



* Reject h(0) at alpha = 0.05, accept it at alpha = 0.01
* Remember we don’t want to make Type II errors (fail to reject h(0) when we should’ve)
* Ex: fail to find effects of new drugs when they actually exist, which is bad b/c research costs a lot of time + money
* So we need to make it more likely we’ll find statistical significance when we should
* Do this by increasing **Power** 🡪 test more people (better indicator of whether drug is effective), give each drug to very similar groups of people (this is why we randomize), test large dosages (see how it’ll affect the subject rather than other factors)
* Higher power can come from:
* Larger samples = remember increasing sample size n decreases SE ( SD / Sqrt(n) )
* Lower Var(w) = more sure of significant differences when distributions aren’t overlapping, b/c if distribution is wide = hard to tell is difference is sig
* Tx’s w/ strong effect sizes = make Tx’s conditions last longer, occur more often, or more intense
* 1-Way ANOVA Assumptions
* Normality 🡪 all populations from which samples are from are normally distributed
* Homogeneity of variance 🡪 data comes from populations that have equal amounts of variability
* Independence of observations 🡪 results from 1 sample won’t affect the others
* *Can* violate these assumptions under *certain conditions*
* Violate normality assumption if sample size is large
* Violate Homogeneity of variance if all sample sizes are similar + ratio of any 2 variances <> 4
* Can use random assignment to conditions to meet independence of observations condition
* Summary
* W/ 3+ samples and we want to know if any 2 are sig diff, we look at both **variability**:
* **between groups** 🡪 how spaced apart sample means are
* find grand mean + each squared deviation from the grand mean for each sample mean and then multiply each of these results by that sample’s sample size
* **w/in groups** 🡪 squared deviation of each value in each sample from their respective sample means + add them up
* Then find the average **sum of squares** for each variability by dividing by respective **dF** (b = # of samples – 1, w = total # of values - # of samples)
* This gives **F**, which if it falls w/in F-crit, reject h(0)
* Can then use **multiple comparisons test** (like **Tukey’s HSD** where if any 2 samples means have a difference greater than this, they’re honestly significantly different) to determine *which* are sig diff
* Can then find the proportion of difference between means that is due to the IV (**eta^2**)