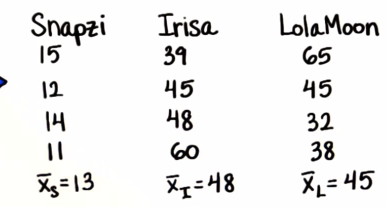
***Udacity Data Analyst Track***

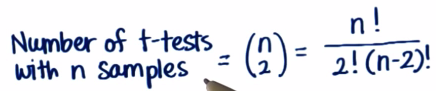
**I. Into to Inferential Stats**

7. 1-Way ANOVA

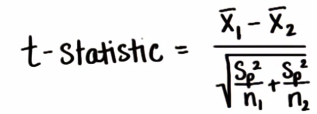
* 3 Brands of clothing w/ random selection of shirts from each w/ the following prices



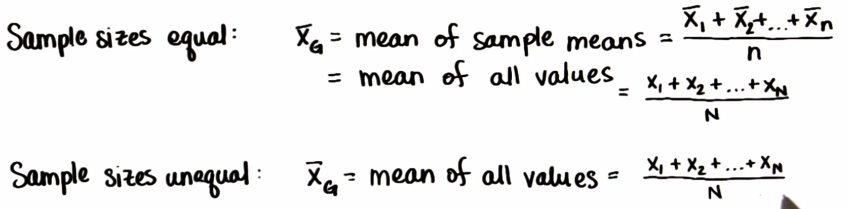
* Could do 3 t-test to statistically figure out if I and L have similar prices or not, or by a simpler method
* Ex: 4 samples 🡪 need 6 t-tests to test for significant differences between 4 independent samples 🡪 A vs. B, A vs. C, A vs D, B vs. C, B vs. D, C vs. D

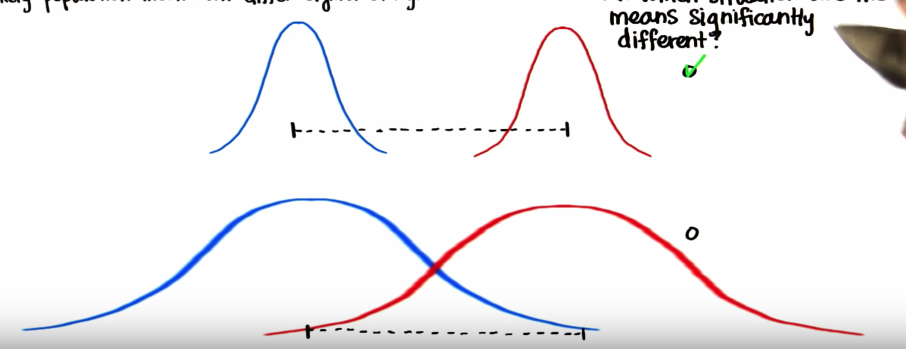
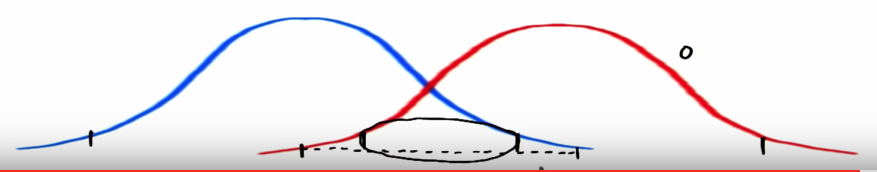


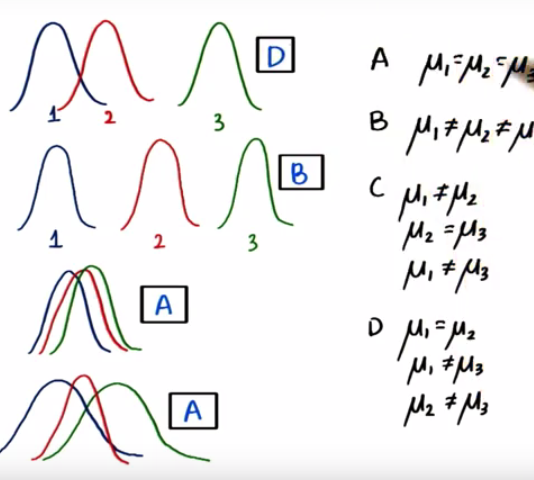
* Don’t want to do a lot of t-tests, but we can use the underlying idea to do it in a simpler way
* T-test 🡪 decision of whether or not 2 samples are significantly different is a function of the distance between their means divided by the variability of each sample (the SE), which is found with the pooled variance

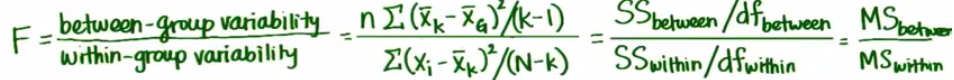


* Can do almost the same thing w/ 3+ samples , since we have distance/variability between all the means, and some kind of error in the denominator
* To get this distance/variability, we want to find the average squared deviation of each sample mean from the *total* mean = **grand mean x(g)** = mean of ALL values in ALL samples
* Since we’re only concerned w/ the variability between means, not the variability between each sample, similar to how we calculated SD
* The grand mean (mean of sample means) will *sometimes* be equal to the mean of all values from all samples, and they are equal when each sample size is the same, since there’s equal weight to each sample in the grand mean



* From the deviation of each sample mean from the mean of means, we can draw the conclusions that the greater the difference between samples means, the more likely the population means will differ, and vice versa = **between-group variability**
* 
* Ex: The means are only significantly different in the top scenario, because we can see their variability w/in each sample (width) is less
* *The more in-sample variability, the less significant differences in sample means are*
* There is an overlap in the possible areas of 95% confidence of where the population means fall for the bottom scenario, wherein the populations these samples came from have the same mean
* 
* This tells us that for comparing 3+ samples, the greater the variability of each individual sample, the more less likely the population means will differ significantly, and vice versa = **w/in-group variability**
* So when comparing a lot of samples, we are extending the idea of the t-test to compare samples by seeing how far a sample means from the grand mean/mean of means (*between-group*) and by looking at the variability of each sample (*w/in group* = impacts if samples are significantly different)
* This whole process is called **Analysis of Variance (ANOVA)**, which can compare as many means as we want w/ just 1 test
* **One-way ANOVA =** one independent variable, or a **factor]**
* H(0): u1 = u2 = u2
* H(a): at least 1 pair of samples is significantly different
* We end up w/ 1 statistic, either big or small
* Small statistic 🡪 *w/in-subject* variability is large relative to *between-subject* variability + NONE of the means are significantly different from one another 🡪 accept h(0)
* Large statistic 🡪 *between-subject* variability is large relative to *w/in-subject* variability + at least ONE pair of means is significantly different from one another 🡪 reject h(0)
* But we don’t know WHICH means are causing the large statistic
* Need an extra steps to see *which* means are different to each other = **multiple comparisons tests** (these are follow-up tests)
* If the variance of an individual sample gets bigger, all else constant, this leans toward h(0) b/c this is w/in-subject variability 🡪 more likely their sample means are similar (one distribution is now wider)
* Opposite goes for between-group variability 🡪 size stays the same, distance between distribution increase = less chance samples means are similar
* Have to construct our statistic such that as w/in-group variability increases, it decreases, and as between-group variability increases, it increases
* Do this by making it a ratio of both variabilities = **F-Ratio** = **BG/WG**
* Numerator = measure of difference between samples, denominator = measure of error



* \*\* Could have multiple H(a)’s 🡪 just need at least 1 mean to be different
* To measure the spread of sample means (between-group), we decided on using the same idea of SD but with the grand mean (each squared deviation from mu(g) and multiply each by the sample size)
* = **(n(k)\*sum[(x(k) – x(g))^2)]/(n-1)** where k = # of samples (k sample means)
* Assuming all samples have the same size 🡪 **(n\*sum[(x(k) – x(g))^2])/(n-1) = (n\*sum[(x(k) – x(g))^2])/(k-1)**
* For w/in-group, take the SSE for each sample from the mean of each sample and divide by dF
* = **sum[(x(i) – x(k))^2]/(n-1) = sum[(x(i) – x(k))^2]/(N-k)**
* dF could be
* n1 + n2 + n3 – 3
* n1 + n2 + n3 – k (where k = # of samples)
* N – k (where N = total # of values in all samples + k = # of samples)
* 4 Tx’ w/ independent samples of participants w/ n = 10 🡪 dF(b) = 4 – 1 = 3, dF(w) = (4n) – n = 36
* F signals whether or not there is a significant difference between any 2 sample means out of k samples
* Can also call it the **[ sum of squares for between groups / dF for between groups ] / [ sum of squares for w/in groups / dF for w/in groups] = [ SS(b)/dF(b) ] / [ SS(w)/dF(w) ] = Mean Square(b) / Mean Square(w)**
* 
* Can see the difference between dF(b) = (k-1) and dF(w) = (N – k) = N – 1, the **total dF, dF(t)**
* 2 = k – 1 🡪 k = 3 24 = N – 3 🡪 N = 27 N – 1 = 26
* In addition to a total dF, we have a **total variation = SS(b) + SS(w) = SS(total) = Sum[(x(i) – x(g)]^2**
* ANOVA partitions the total variation into between and w/in groups variations b/c differences in the treatment/outcome/dependent variable are due to both between-group and w/in (between individuals)-group differences
* So, only, *some* variation can be explained by knowing which group a subject is in, and the rest is unexplained
* Unlike z and t tests, the F-statistic is NOT symmetrical, and is positively/right-skewed and peaks at 1



* This is b/c if there are NO differences in the population means (BG variability expected to be 0), then the mean of each sample will still differ by chance
* Since the difference is due to chance in the same way that each subject in each sample differs by chance (measured by WG variability), BG and WG will be equal so F = 1
* So, if h(0) is true, then BG variability is expected to be 0, so F is expected to be 1
* Also, F-tests are NON-DIRECTIONAL 🡪 only know that there IS a difference, not in which way
* Critical region is in upper tail, and we choose it the same based on alpha and then find F and see if it lies in the critical region
* Largest F’s 🡪 Large BG and small WG
* If so, at least 2 population means will be significantly different
* Also, note that sources of w/in-group variability include individual differences, experimental error, + ambiguous survey questions, while treatment effects would only affect between-group affects.