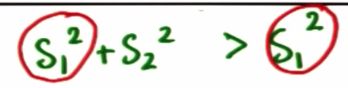
***Udacity Data Analyst Track***

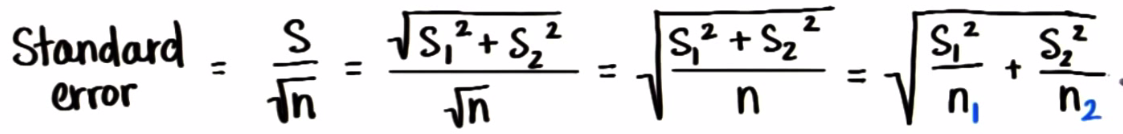
**I. Into to Inferential Stats**

6. T-tests Part 3

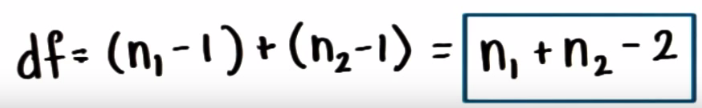
* Before, we’ve learned about **dependent samples (within-subject designs)** = repeated measures
* Could be giving a person **2 conditions**/treatments + seeing how they react (control vs. Tx or 2 types of Tx’s)
* Could be Longitudinal 🡪 measure a variable at 2 points in time to see if it changes
* pre/post test 🡪 measurement of variable before + after treatment
* Advantages of dependent samples:
* Controls for individual differences (same in each treatment given for an individual)
* Can use fewer subjects, cost-effect, less time-consuming, + generally less expensive
* Disadvantages of dependent samples:
* **Carry-over effects** 🡪 ex: teach students 1 way of doing math in 1 lesson and another way in a 2nd lesson, they will most likely be better at math anyway in the 2nd lesson, regardless of teaching method
* **Don’t know if results are due to an effective treatment or from doing more math**
* i.e. second measurement can be affected by the 1st Tx
* So order of Tx’s may influence results (1st pill has interaction w/ 2nd pill)
* **Independent samples (between-subject designs)**
* Advantages of dependent samples:
* No Carry-over effects
* i.e. second measurement is not affected by the 1st Tx
* Can give 1 Tx to 1 group and another to another group + not worry about interactions
* Each subject gets only 1 Tx
* So order of Tx’s may influence results (1st pill has interaction w/ 2nd pill)
* Disadvantages of dependent samples:
* Can’t control for individual differences
* Need more subjects (larger n) + to randomize the 2 groups taking the 2 Tx’s to control for this as best as possible
* Makes it less cost-effect, more time-consuming, + generally more expensive
* **Experimental study** 🡪 give Tx to subjects
* **Observational study** 🡪 observe characteristics of 2 different populations + compare
* Independent sample studies use the *same* h(0) and h(a) technique + *same* methods on how we make a statistical decision (p < alpha or not) as dependent sample studies
* But they have *different* SE’s and a *different*  t statistic, b/c now we have 2 sample sizes and 2 samples SD’s
* If we subtract normally distributed data from another normal distribution, we end w/ a new set of data 🡪 N(mu1,S1) – N(mu2,S2) = N(u1-u2, Sqrt(S1^2 + S2^2)
* Roughly the same for samples
* This new SD is greater than each of the individual data sets (b/c it’s a sum)



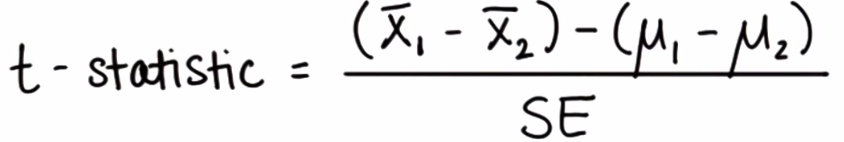
* Intuitive b/c if we subtract a data set from another, there’s going to be even more error than exists in each individual dataset 🡪 *wider distribution*
* W/ independent samples, we are analyzing the difference between those 2 means + the *new* SE is therefore **SE = Sqrt(s1^2 + s2^2)/Sqrt(n) = Sqrt((s1^2+s2^2)/n) = Sqrt(s1^2/n + s2^2/n)**
* If the samples sizes are different, use n1 and n2 🡪 **Sqrt(s1^2/n1 + s2^2/n2)**



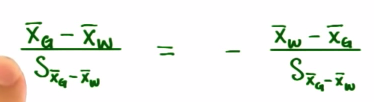
* ***NOTE***: we are squaring the SD 🡪 we are calculating the **variance**!
* W/ paired data/dependent samples, we could calculate the difference between each value for each subject and then calculate SD (only 1 S), but now we have 2 samples that could differ in size
* And now the dF changes 🡪 subtract 1 from *each* sample size + add them



* If not using software, a more conservative approach = take the smaller of n1 - 1 and n2 - 1
* New t-statistic just replaces x w/ the sample means difference and mu with the population mean differences



* Valid in 2 ways:



* If we want to eat but not pay a lot, we look at average meal prices for 18 restaurants in Gettysburg and 14 in Wilma
* h(0) = pr(g) = pr(w) h(a) = pr(g) <> pr(w)
* To compare these samples, we need: the sample averages, sample SD’s, and sample sizes
* Check Excel
* Dermatologist has developed drug A to treat acne, tested on 6 people, another Dermatologist developed drub B and tested in on 5 people
* Check Excel
* Our SE calculation so far has assumed the sample sizes are similar/approximately the same size
* Must correct for this by **pooling the variances**, which will change the SE a bit
* S(p)^2 = average of 2 sample variances that corrects for different samples sizes
* W/ 1 sample, the variance = sum of squared deviations / dF 🡪 **SS/dF = sum[(x(i) – x)^2]/(n-1)**
* w/ pooled variances, we’re doing almost the same thing 🡪 **s(p)^2 = (SS1 + SS2) / (df1 + df2)**
* **Corrected SE = Sqrt(s(p)^2/n1 + sp(2)^2/n2)**
* **Corrected t 🡪 mean difference – expected mean difference / corrected SE**
* **(x – y) – (mu(x) – mu(y)) / SE(xy)**
* if h(0) says no expected difference 🡪 **(x – y) – (0) / SE(xy)**
* In general, use the pooled variance b/c it corrects for sample sizes + the calculation of t is basically the same 🡪**(x – y) / SE(xy)** , when we expect both populations would be the same (difference of 0)
* If we expect the difference between populations to be 10 (**observed difference)**, do [**(x – y) – 10] / SE(xy)**
* Our typical h(0) is no expected difference, so 0 is used most often
* T-test assumptions for using pooled variance:
* X + Y = random samples from 2 *independent* populations
* Both source populations for X and Y should be approximately normal (less important when n is very large [ > 30]
* Sample data can be used to estimate population variances
* Populations variances should be roughly equal so that we can used the pooled variance as an estimate of both of them