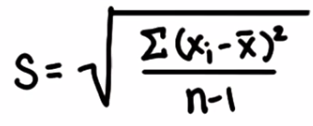
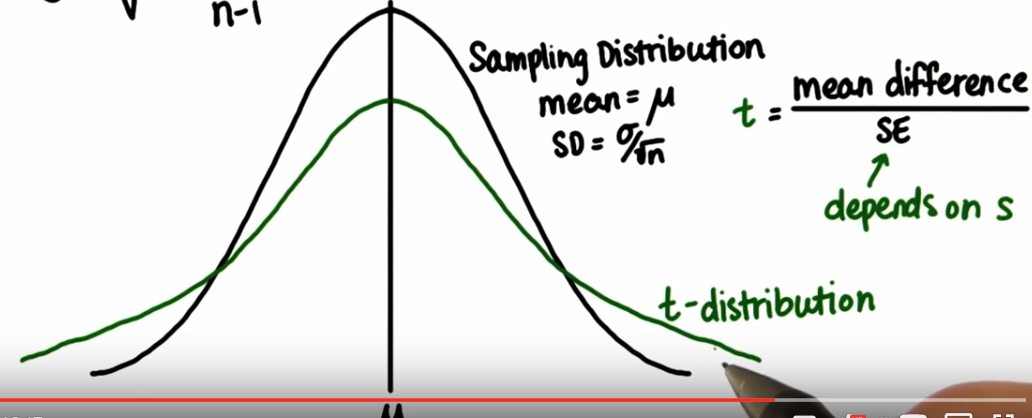
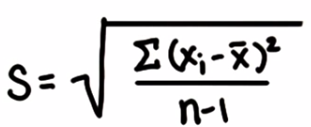
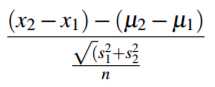
***Udacity Data Analyst Track***

**I. Into to Inferential Stats**

4. T-tests Part 1

* Much of the time we actually do NOT know the population mean and SD, and only have samples
* But we can use these samples to figure out how different a sample mean is from the population and from other sample means (can be **dependent** or **independent**)
* When working w/ samples, we need to *estimate* the pop SD using sigma w/ **Bessel’s correction = sqrt(sum(xi – x)^2/(n-1)) =** 
* Before, we found where a sample mean was in the distribution of sample means (whose shape we know from the population parameters)
* For any sample mean, we could find where it was in the distribution by **standardizing** (finding z-score = mean difference / SE)
* But SE depends on the sample in this case since we don’t know sigma 🡪 end up w/ a distribution more prone to error (**t-distribution**) = more spread out + thicker @ tails
*  
* Just like before when larger sample sizes = skinnier distributions 🡪 as n increases here, t-distribution gets skinner, t-distribution approaches a normal distribution, SE decreases (larger sample size = less error), and SE approaches SD
* The **t-Test** is best to use when we DON’T know the population SD + we instead use sample SD (SE)
* The **t-Test statistic** can be computed similarly to the **z-score** 🡪 **t = x - µ / sigma/sqrt(n) = x - µ / SE**
* We also have to compute **degrees of freedom (df)** for the sample 🡪 **df = n - 1**
* How many numbers we are free to choose before we are forced to choose (4 numbers to add up to 10 🡪 can pick 1st 3, 4th # is forced to sum up to 10, or sudoku, where we pick #’s in a row, but the last one must be forced to be the remaining # in a row/column)
* **dF** = # of pieces of info that can be freely varied w/out violating any given restrictions/@ of independent pieces of info available to estimate another piece of info
* Only n-1 pieces of info are available after we know the mean
* The sum of x1 to xn values *must* = mean(x) \*n, so we have the freedom to choose n-1 values for x, but the last one must make it such that the sum of all x’s = mean(x)\*n
* As dF increases the t-distribution better approximates the normal distribution
* **n – 1 = effective sample size 🡪** n - 1 values may vary, but the last value must make the sum = mean(x)
* Like the Z-score, we can use the t-table to get the proportion below or between a specific value
* T-table tells us the **t critical values**
* When t-statistic is far from 0 in either direction (sample mean is far from pop mean), reject the NULL
* T statistic = (smpl mean – pop mean) / sample error 🡪 (smpl mean – pop mean) / (smpl SD/sqrt(n))
* When comparing sample mean to initial population mean, center the t-distribution at pop mean and find where the sample mean lies on the distribution
* Further away, more significant chance it comes from a population w/ a significantly different intervention population mean (
* larger s mean = new pop mean > old pop mean, smaller s mean = new pop mean < old pop mean
* 1 sample t-test:
* Null = pop mean = old pop mean (mu = mu(0))
* Alternative 🡪 mu < m(0), or mu > mu(0), or just mu =/= mu(0)
* t numerator = point estimate for population mean mu/sample mean minus mu(0) (difference between)
* t denominator = measures difference between mu and mu(0) that we expect by chance
* To increase t (the t statistic), we can have a larger difference between x and mu(0) or a larger sample size
* Larger t = higher on distribution = lower probability of obtaining that t value
* T-tests are also great for testing 2 sample means (**paired t-tests**) where we modify the formula to become: **(x2 - x1) - (µ2 - µ1) / (sqrt(s1^2 + s2^2)/n) =**  
* **Cohen’s d** measures the **effect size** of the strength of a phenomenon + gives us the distance between means in standardized units + is computed by 🡪 **d = (x1 - x2) / s**
* where s = sqrt((n1 - 1)\*s1^2 + (n2 - 1)\*s2^2)/(n1 + n2 – 2) =
* Ex: Pizza company A wants to know if they deliver faster than Company B

|  |  |
| --- | --- |
| A | B |
| 20.4 | 20.2 |
| 24.2 | 16.9 |
| 15.4 | 18.5 |
| 21.4 | 17.3 |
| 20.2 | 20.5 |
| 18.5 |  |
| 21.5 |  |

* Use Cohen’s d to measure the effect size between the two times.