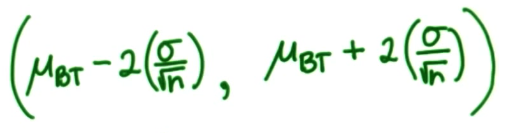
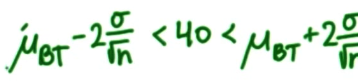
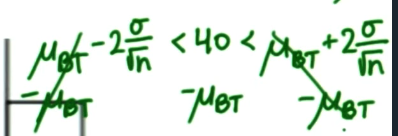
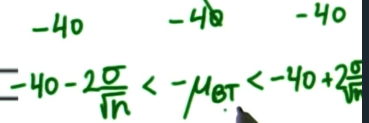
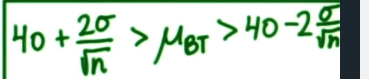
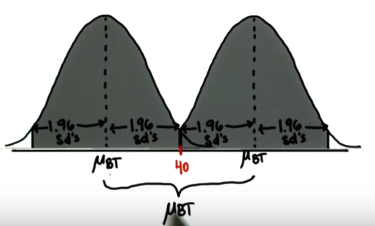
***Udacity Data Analyst Track***

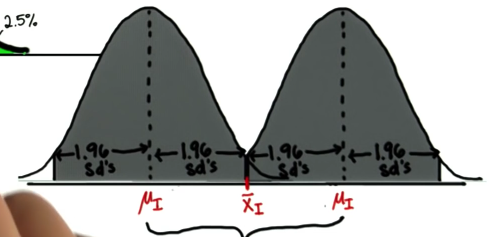
**I. Into to Descriptive Stats**

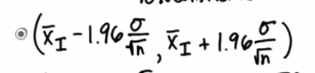
2. Estimation

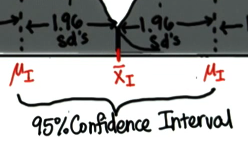
* John Tucket 🡪 *an approximate answer to the right problem is worth much more than the exact right answer to the approximate problem*
* **Margin of error =** amount of error predicted when estimating population parameters from sample statistics 🡪 computed as **Z\*** · **sigma/sqrt(n)**
* Z\* = the **critical z-score** for the level of confidence.
* **Confidence level** = estimate of the % of all possible sample means that fall w/in a margin of error of our estimate.
* “we are some % sure the *true* population parameter falls w/in a specific range”
* **Confidence Interval =** a range of values in which we suspect the population parameter lies between.
* To compute the confidence interval we use the formula: x¯ +/- Z\* · sigma/sqrt(n)
* Basically the margin of error times the sample mean
* This gives us an upper (+) and lower bound (-) that captures our population mean.
* **Critical Z-Score** = used to define a critical region for our confidence interval.
* Observations beyond this critical region = observations so extreme that they were very unlikely to have just happened by chance.
* Klout review 🡪 pop = 1048, mean = 37.72, Sigma = 16.04
* From the CLT, if we took all possible samples of the same size + found the mean of each sample and graphed the distribution (**sampling distribution**) of those sample means, we get a normal curve w/ mean = sigma and the sampling distribution SD/**Standard Error (SE)** = sigma / Sqrt(n)
* If we took a sample of size 35 of users who used Bieber Tweeter and found the mean = 40 (a **point estimate**), then if *everyone* *in the population* started using Bieber Tweeter we’d expect our best *GUESS* of the pop. mean to be = 40
* There’s some range around 40 wherein the TRUE population mean would be in this case
* Approximately *95% of sample means fall between* ***2(Sigma)/sqrt(n))*** [ OR 2 times the SE ] in this case, and 68% are w/in (Sigma)/sqrt(n)) [ OR 1 SE ]
* Here, for the 95% example, 2(Sigma)/sqrt(n))= **margin of error**
* If everyone started using Bieber Tweeter, there would be a new population distribution, but we don’t’ know what it would be
* But we *do* know a sample of size 35 has mean = 40, and if 95% of sample means are w/in 2 SE’s of the population mean, then this sample mean value of 40 has a 95% chance of being w/in 2 SE’s of the *new* population distribution’s mean
* We have a new best guess for the Bieber Tweeter population mean that we get from the sample mean of only Bieber Tweeter users, called **u(bt)**
* And we know 95% of sample means of people who use BT would fall w/in some interval
* 
* And it’s *very* likely this sample mean = 40 is w/in this interval (is one of those 95% of sample means)
*  🡪 
*  🡺 
* 
* W/ u(bt) = 40, we’d have an interval of:
* (40 - (2\*16.04)/Sqrt(35)) < u(bt) < (40 - (2\*16.04)/Sqrt(35)) = **34.57749 < u(bt) < 45.42251**
* This is our **confidence interval (CI),** more specifically our **95% CI, +** our distribution can be anywhere from being centered at u = 34.58 to being centered at u = 45.42, where a value of 40 is still w/in that 95% range for both cases (w/in 2 SE’s)
* Note, we are sure *APPROXIMATELY 95%* of sample means fall w/in this interval, so 2\*Sigma/Sqrt(n) or 2 \* SE is *APPROXIMATELY our margin of error*
* In a normal distribution w/ mean u, approximately 95% of data are w/in 2 SD, but the *exact # of SD’s* that bounds 95% of the data can be found from the z-table and is **-1.96 (0.0250) + 1.96 (0.9750)**
* So w/ a NORMAL sampling distribution, *EXACTLY* 95% of samples means are w/in 1.96 SE’s from the population mean
* Calculate 95% CIfor the mean of all BT users using **point estimate** of 40
* SE = (16.04)/Sqrt(35) = **2.711254849** 🡪 SE\*1.96 = **5.314059504**
* {40 - 5.314059504, 40 + 5.314059504} == **{34.68594, 45.31406}**
* *So exactly 95% of samples means in this distribution are between* ***{****34.68594****,*** *45.31406}* = **interval estimate**



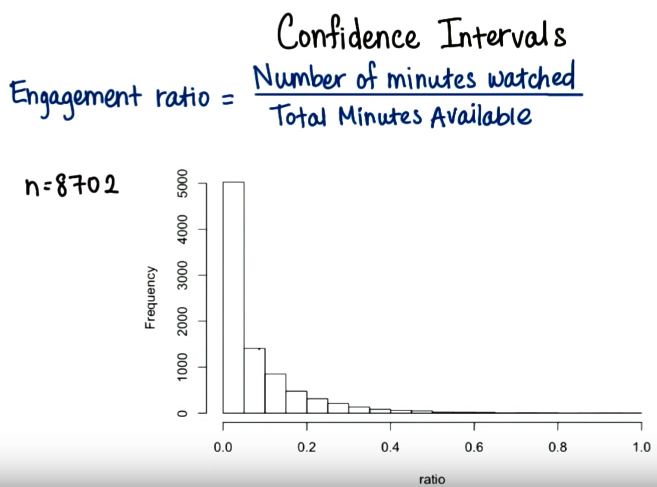
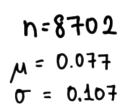
* Say we have some sample mean, x(t), and after some intervention, we want to find the population mean afterward (i.e. everyone use BT)
* Pop. mean after intervention can be anywhere in:



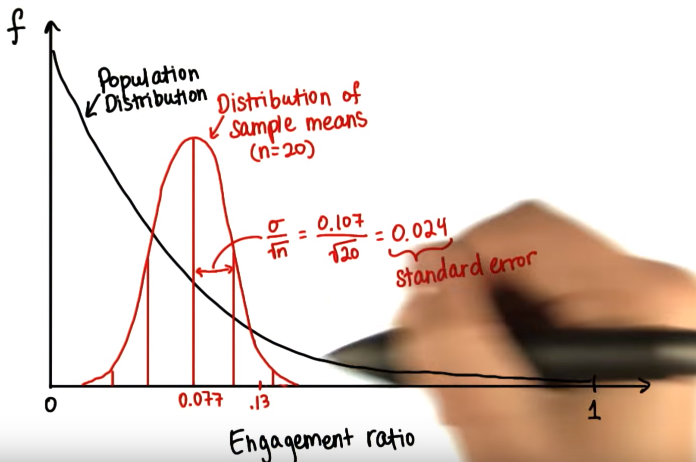
* So long as x(t) is w/in 1.96 SE’s of the mean
* Point estimate here is x(t) (from a sample of size n of just those who got the intervention and take the mean)
* Interval estimate: 



* Can’t compute yet b/c we don’t the population mean after the intervention
* Back to BT, we know our sample mean/point estimate = 40, the population mean ranges from {34.68594, 45.31406}.
* We know the *original* pop mean = 37.72, which is well w/in this interval
* So, a sample of size 35 w/ mean 40 *could’ve very well been selected by random chance*
* What about w/ sample size = 250?
* SE = (16.04)/Sqrt(250) = 1.014459🡪 SE\*1.96 = **1.988339**
* {40 - 1.988339, 40 + 1.988339 } == **{38.01166, 41.98834}**
* So, standard error is smaller, so 1 SE away from the mean is a smaller range, so our CI is smaller
* *Larger sample size = more precise estimates via smaller CI’s for the true population mean (parameter) after intervention due to the smaller SE (SD of sampling distribution)*
* i.e. shrink an interval in which a sample mean might feasibly lie
* Notice new CI doesn’t contain pop. mean = 37.72, which means its very unlikely a random sample of size 250 w/ mean 40 was selected by chance *= evidence of an effect of a treatment*
* What about a 98% CI to be more sure the new population mean is w/in some interval 🡪 2.33 SD’s
* SE = (16.04)/Sqrt(250) = 1.014459🡪 SE\*2.33 = **2.363688709**
* {40 - 2.363688709, 40 + 2.363688709} == **{37.63631129, 42.36368871}**
* The CI is now wider, since we need a wider range to be sure *APPROXIMATELY 98%* of sample means fall w/in some interval, such that 1% of the data is above it or below it, so we’re more sure the intervention parameter falls w/in this range
* +/- 2.33 = **critical values of z** for 98% confidence, +/- 1.96 = critical values of z for 95% confidence
* Ex: Engagement Ratio (a construct to be measured)

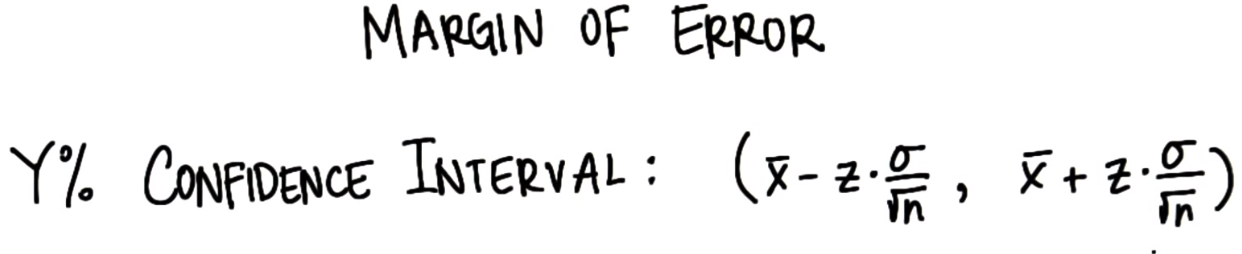
* Intervention sample (n=20) mean (point estimate = 0.13) 🡪 SE = 0.107/sqrt(20) = 0.023973
* Want to know how many SE’s away from the population mean our point estimate is



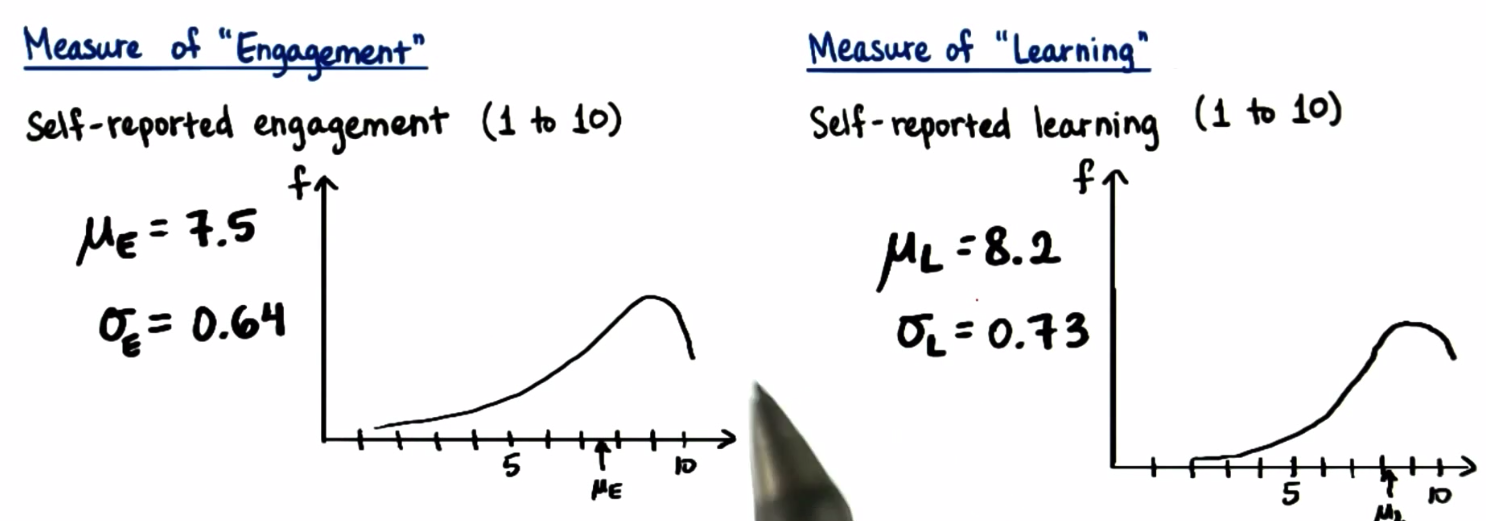
* 95%
* SE = (16.04)/Sqrt(250) = 1.014459🡪 SE\*2.33 = **2.363688709**
* {40 - 2.363688709, 40 + 2.363688709} == **{37.63631129, 42.36368871}**



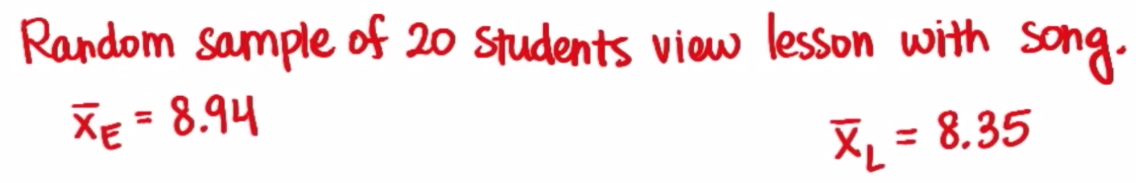




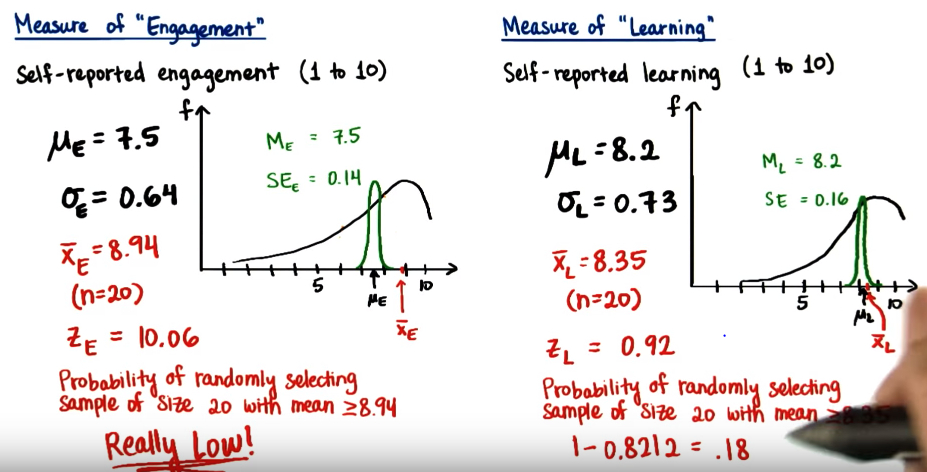
* **Margin of error** 🡪 distance from sample means on either side ( = ½ of the width of the CI)
* **Ex:** Different way of measuring engagement construct



* Negatively skewed distributions 🡪 want to increase those means
* Give some sample an intervention of a song to help engage and learn



* Does this song lead to higher engagement + learning? 🡪 need to find where each sample mean falls on the distribution of sample means for their respective populations



* So the song seems to have had an effect on *engagement*, and NOT on learning (18.5% chance of selecting a sample w/ mean = 8.35, but 0% chance of selecting a sample w/ mean = 9.84)
* Still not sure it CAUSED it 🡪 need experiment for that
* Problem w/ point estimates 🡪 don’t account for sampling error
* 98% CI has larger range of values 🡪 increasing sample size decreases range, increasing SD increases range
* Find a confidence interval for the distribution of pizza delivery times.
* Company A
* 20.4
* 24.2
* 15.4
* 21.4
* 20.2
* 18.5
* 21.5