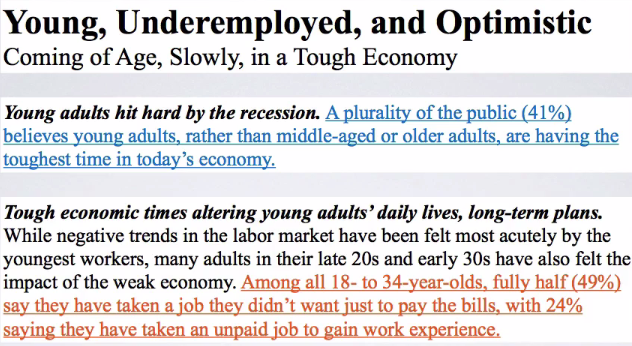
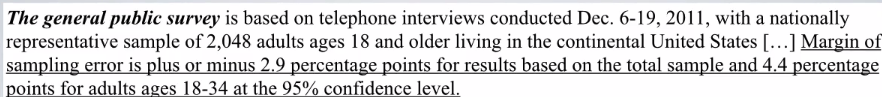
***COURSERA: STATS W/ R SPECIALIZATION***

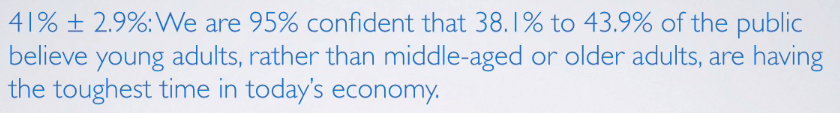
***COURSE 2 - Inference***

**WEEK 1- Central Limit Theorem and Confidence Interval**

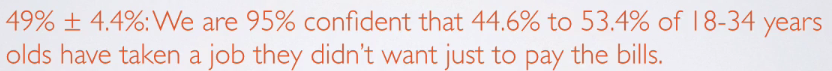
***4.1 CLT and Sampling***

**Intro**

* 
* 
* Remember, the study had estimated 41% of the public believes young adults, rather than middle aged or older adults, are having the toughest time in today's economy.



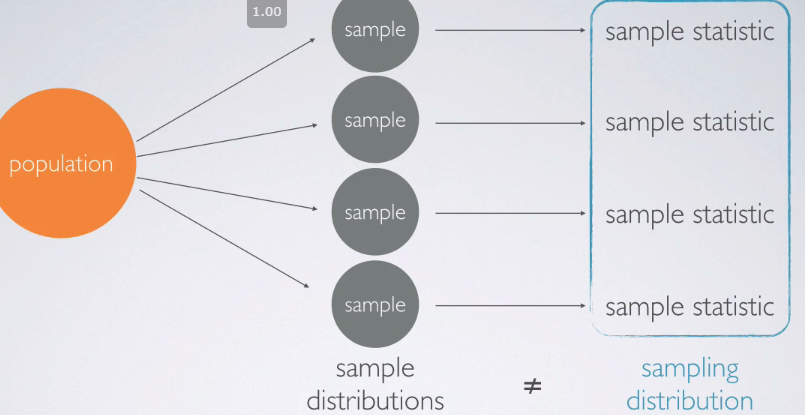
* We were also told that 49% of the public had taken a job they didn't want just to pay the bills.



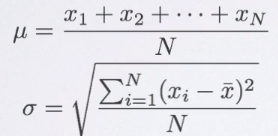
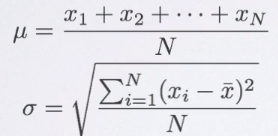
* The 41 + 49% we have on hand come from **sample statistics**, but we're often interested in **population parameters**.
* Since compete populations are difficult/impossible to collect data on (i.e. collecting data from the entire US population for this study), use sample statistics as **point estimates** for unknown population parameters of interest.
* But samples statistics vary from sample to sample = other random samples of Americans would yield slightly different estimates
* *Quantifying* how sample statistics *vary* provides a way to estimate **margin of error** associated w/ a point estimate.
* Discussion on **sampling variability** (how estimates vary from 1 sample to another) is important.
* The **CLT** describes shapes, centers, + spreads of sampling distributions when certain conditions about the population, as well as the sampling scheme, are met

**Sampling Variability and CLT**

* Say we have a population of interest + a sample from it that we calculate sample stats from it, then do the same for more samples
* Each sample will have its own **sample distribution,** + each observation in these distributions is a randomly sampled unit from the population
* The sampling statistics from each sample *also* make a sampling distribution, the **SAMPLING distribution** where each observation is a sample statistic and *not* a unit from the population



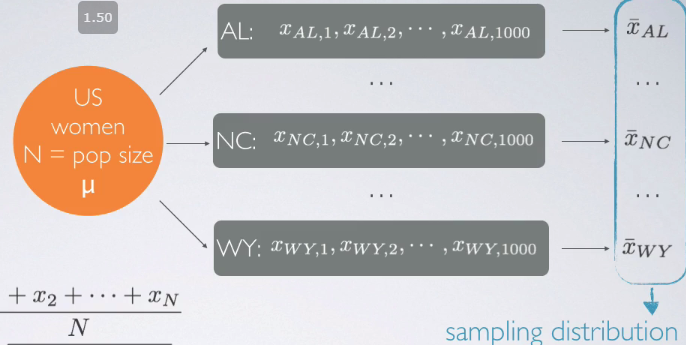
* Ex: Average height of US women 🡪 population of interest = US women w/ population size N, parameter = average height of ALL US women, μ
* We could calculate population mean + population SD **σ** (which wouldn’t be small since heights vary so much)

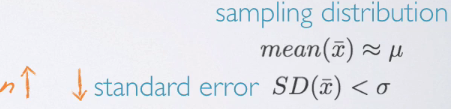
* Sample 1K women from each state, expecting the distribution of sample means from these samples to be less variable than the population
* Ex: Samples from Alabama where X is the observation and the # subscript is the i-th observation



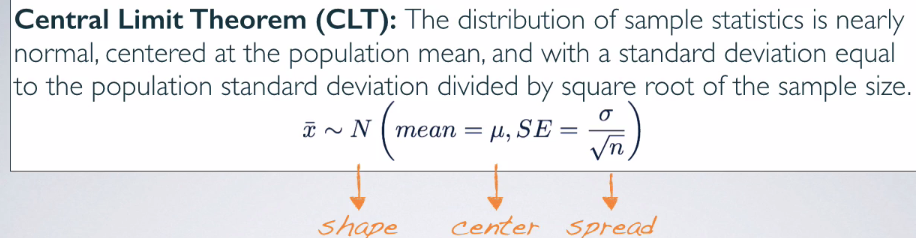
* Then find the mean for each state sample, X\_.
* All these state sample means = sampling distribution



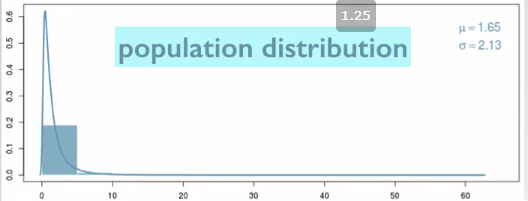
* The mean of the SAMPLING distribution *should* be close to the estimated population mean and the SD of the sampling distribution (standard error, SE, of the means) *should* be smaller than the population’s
* As sample size n increases, SE would decrease, as n is the denominator in its calculation



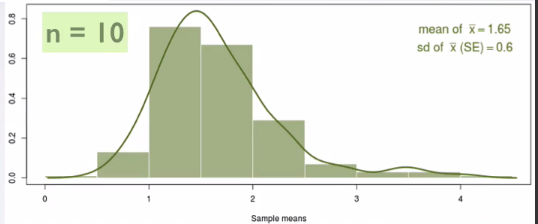
* Therefore expect a skinner distribution when plotted (b/c less variable from the mean)
* While observations can be very variable, it’s unlikely that sample means will be very variable
* As n increases, we’d expect sample means to be more consistent = less variable = lower SE



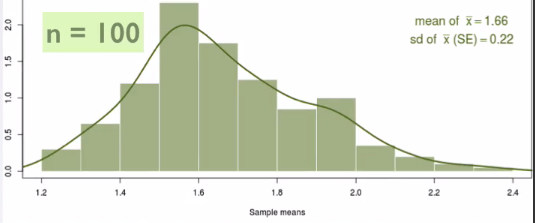
* Or, “The sampling distribution of the mean/distribution of sample means from many samples is nearly normal centered at the population mean, w/ standard error = the population standard deviation divided by the square root of the sample size.
* “Central” Limit theorem 🡺 central to much of statistical inference theorem
* CLT tells us about the shape, center, + spread of a sampling distribution
* *Often population SD is unknown, so we’d use the sample SD,* ***s****, to calculate SE*
* 2 Conditions For CLT To Apply
* **Independence** = Samples observations must be independent
* Very difficult to verify, but it is more likely w/ random sampling/assignment depending on
* Observational study = *sampling* from population randomly
* Experiment = randomly *assigning* experimental units to various treatments.
* Also more likely if, when sampling *w/out replacement*, sample size n < 10% of the population
* Say you live in a small town w/ population = 1K including your family + extended family
* I'm doing research on some genetic application + want to randomly sample some individuals from your town, say of size 10.
* If randomly sampling 10 people out of 1000 + you’re included in the sample, it's going to be quite unlikely your parents are also included in that sample as well b/c we're only grabbing 10 out of a population of 1000.
* But if I sampled 500 people, I have 499 other chances to get somebody from your family in the sample as well, which is more likely
* You + a family member are NOT genetically independent, b/c observations in the population itself are not independent of each other often.
* Therefore if we grab a very big portion of the population to be in a sample, it's going to be very difficult to make sure sampled individuals are independent of each other.
* That's why, while we like large samples, also want to keep sample size somewhat proportional to a population.
* The good rule of thumb usually, if sampling w/out replacement = don't grab more than 10% of a population to be in a sample.
* When sampling *WITH replacement* (not common in surveys = don't need your responses again), the probability of sampling you vs. somebody in your family stays consistent throughout all trials + we wouldn't need to worry about the 10% condition here
* But again, in realistic survey sampling situations, sample w/out replacement + like large samples, but also don’t want samples to be much larger than 10% of a population.
* So we love large samples but we don't exactly want them to be very large.
* **Sample size/skew** 🡪 either population distribution = normal or if skewed/we have no idea what it looks like, the sample size must be large.
* According to the CLT, if the population distribution is normal, the sampling distribution will also be nearly normal, regardless of sample size.
* If population distribution is not normal, then the more skewed a population distribution is, + the larger sample size we need for the CLT to apply.
* For *moderately* skewed distributions, n > 30 is a widely used rule of thumb to make use of
* The distribution of a population is also very difficult to verify b/c we often do not know what a population looks like.
* That's why we're doing this investigation in the 1st place
* We can check it w/ sample data, assuming it mirrors the population
* If you make a plot (histogram, boxplot, normal probability plot) of a sample distribution + it looks nearly normal, might be fairly certain the parent population this distribution is coming from is nearly normal as well.
* Say we have skewed population distribution



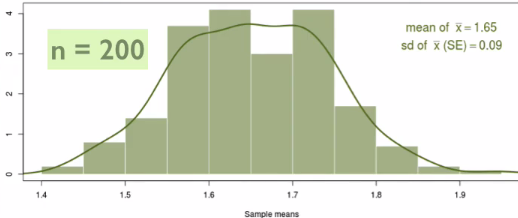
* When sample size = small, the sample means will be quite variable + the shape of their distribution will mimic the population’s.



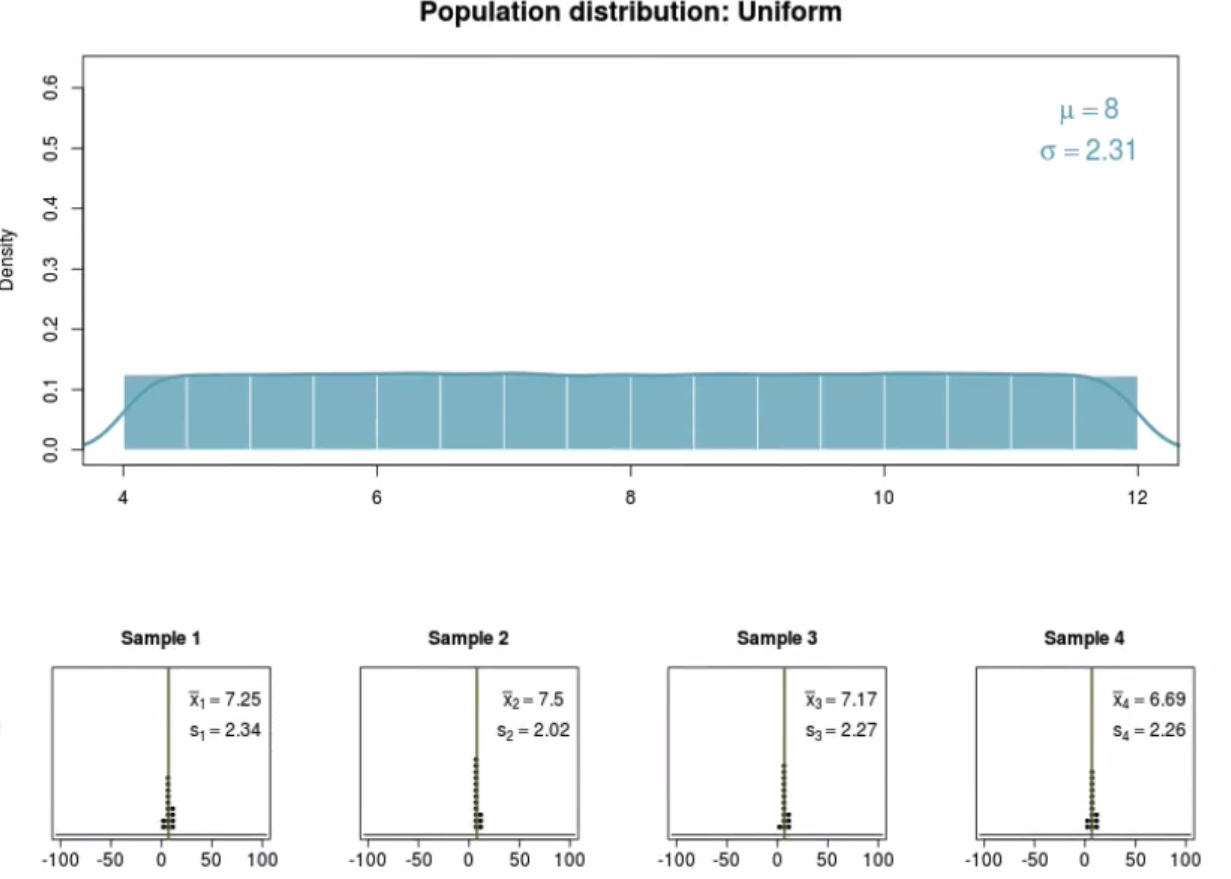
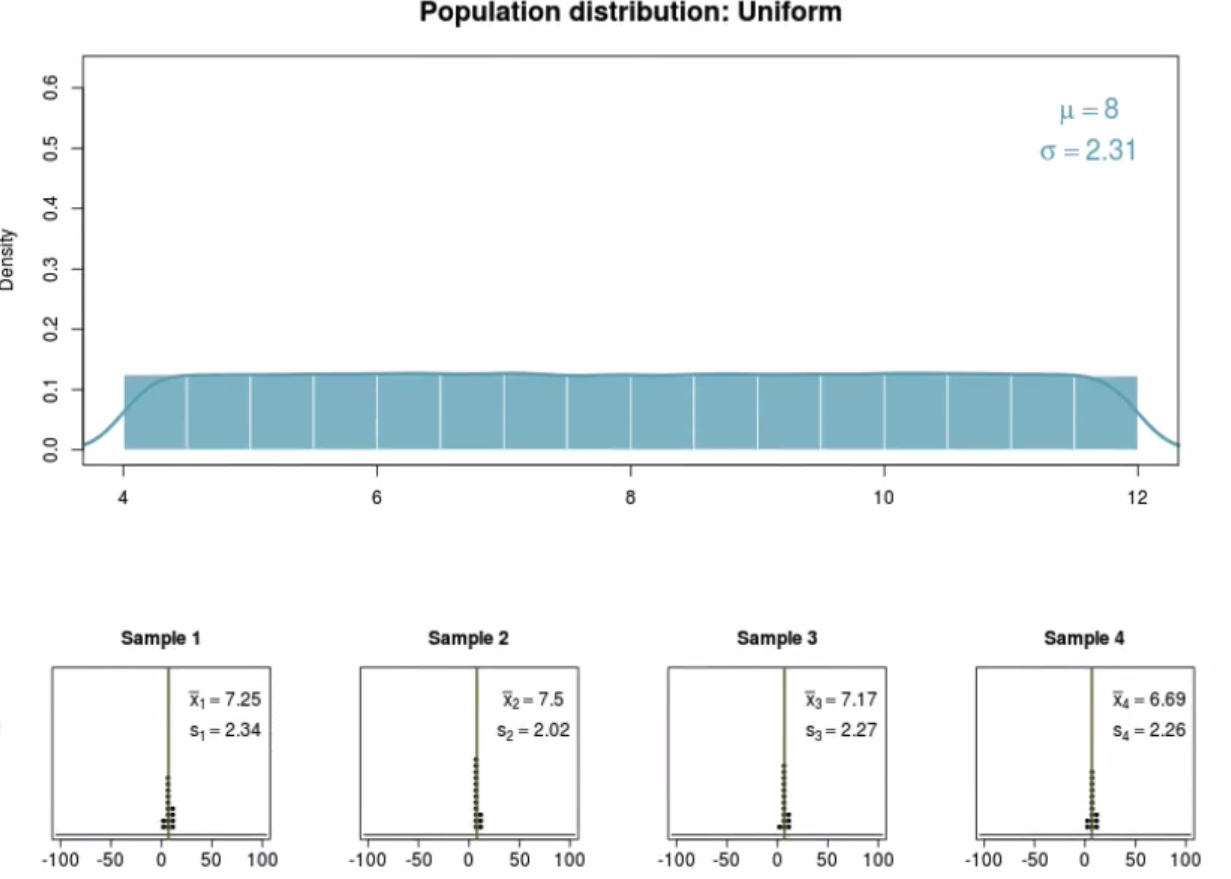
* Increasing sample size = decreases standard error + the distribution starts to condense around the mean + starts looking more unimodal and symmetric.

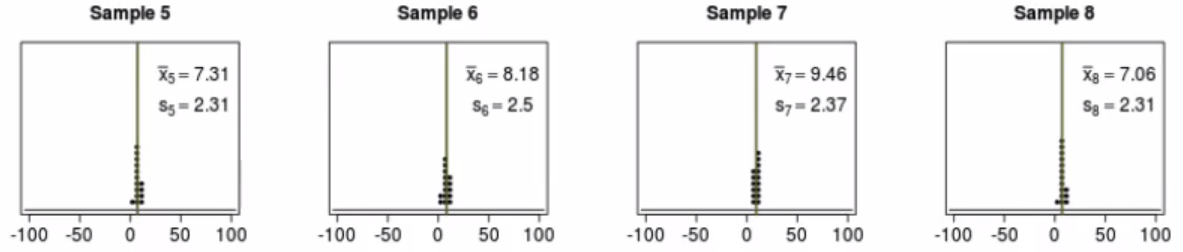


* W/ quite large samples, we can actually overcome the effect of the parent distribution + the CLT kicks in 🡪 sampling distribution starts to resemble a closely normal distribution.

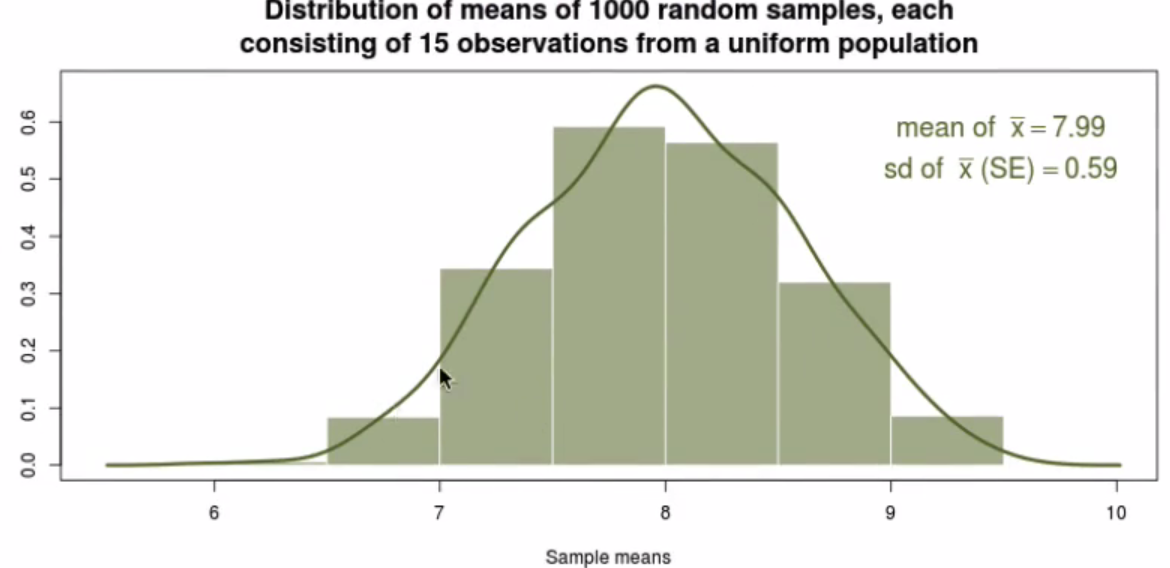


* *Why are we obsessed w/ having nearly normal sampling distributions?*
* Because we've learned earlier that **once you have a normal distribution, calculating probabilities which will later serve as P values in hypothesis tests are relatively simple**.
* So, having a nearly normal sampling distribution that relies on the CLT opens up a bunch of doors for us for doing statistical inference using CI’s + hypothesis test using **normal distribution theory**.
* What happens if the population distribution is NOT nearly normal.
* Ex: Uniform population distribution w/ samples of size n = 15 from this distribution.

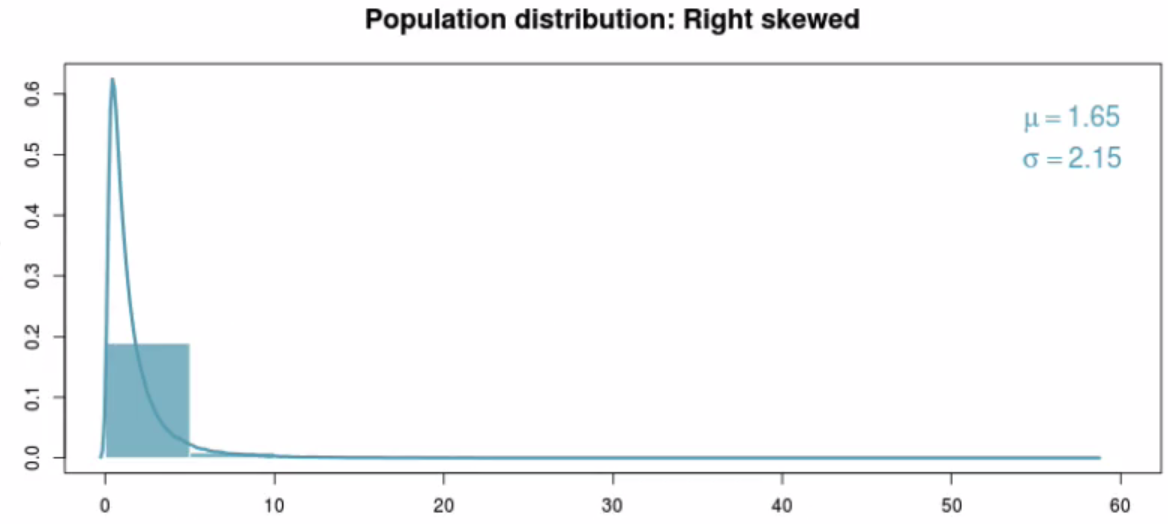


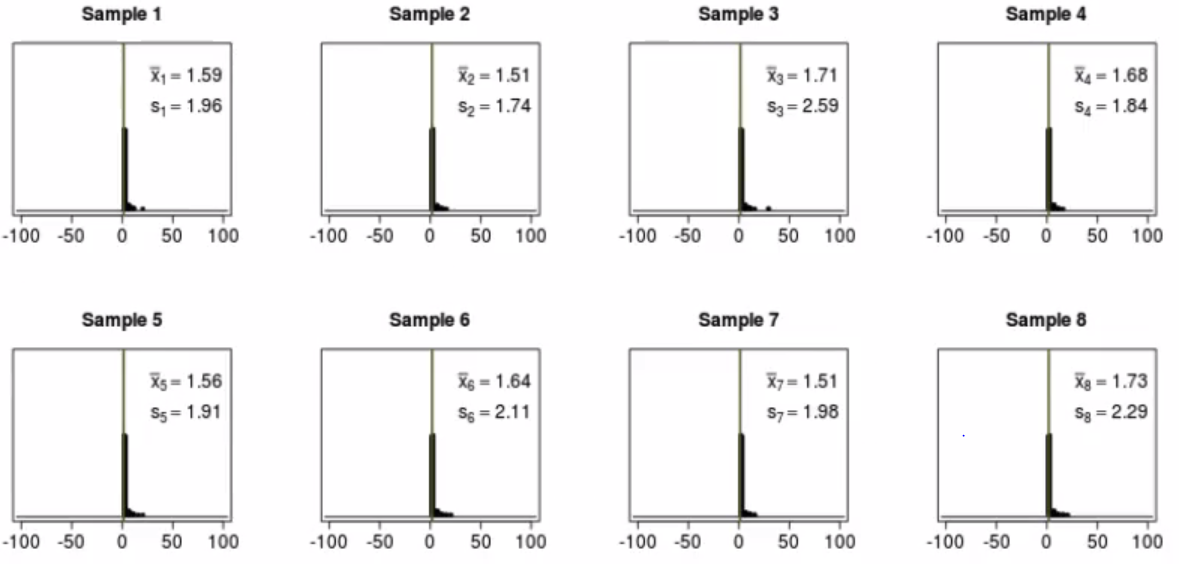


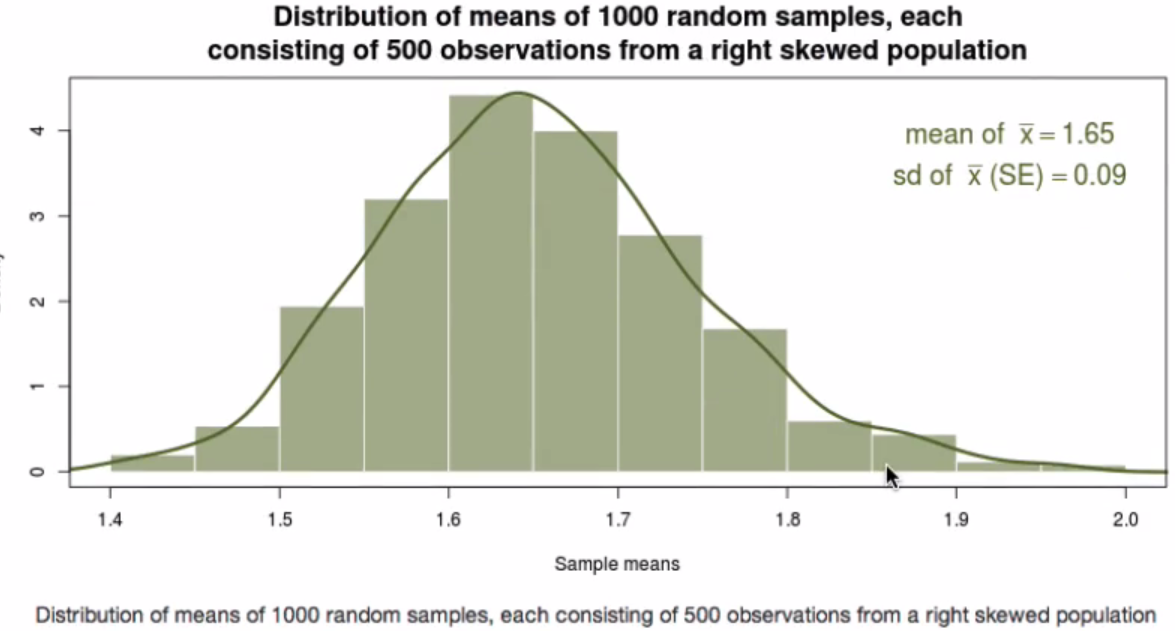
* The center of these samples = somewhere close to the population mean.
* W/ 1000 samples, the sampling distribution looks fairly unimodal + symmetric.

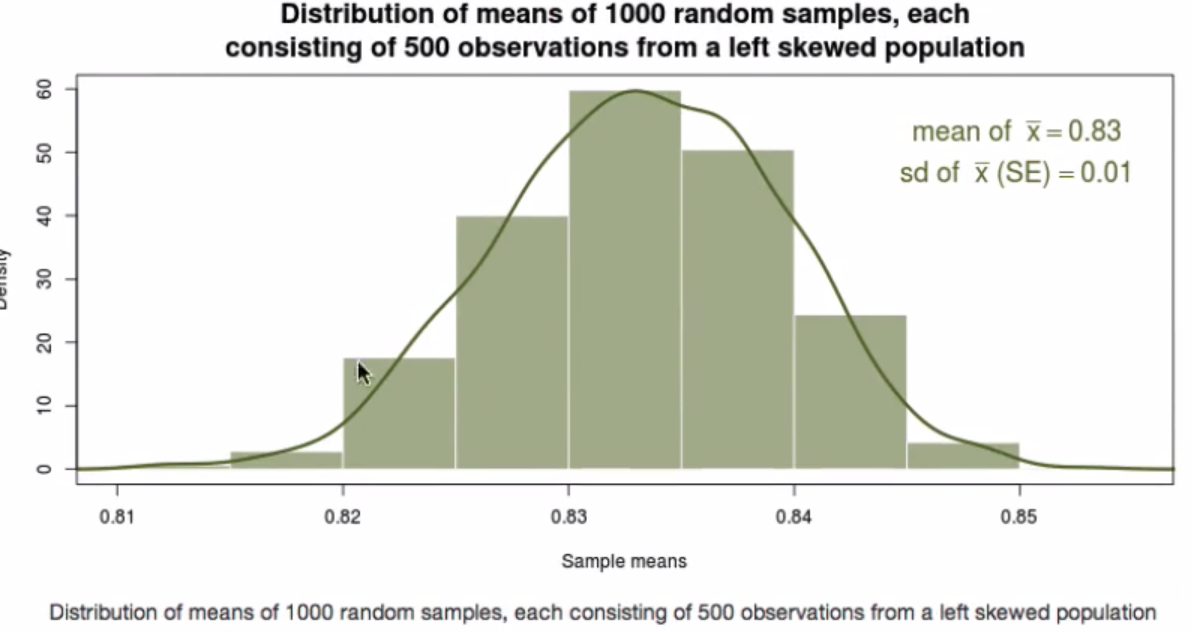


* Center of distribution (7.99) = very close to population distribution mean (8)
* Variability of sampling distribution (SE\_x = .59) = actually much lower than population distribution (SD = 2.31)
* Population distribution that's right skewed w/ samples of size 500

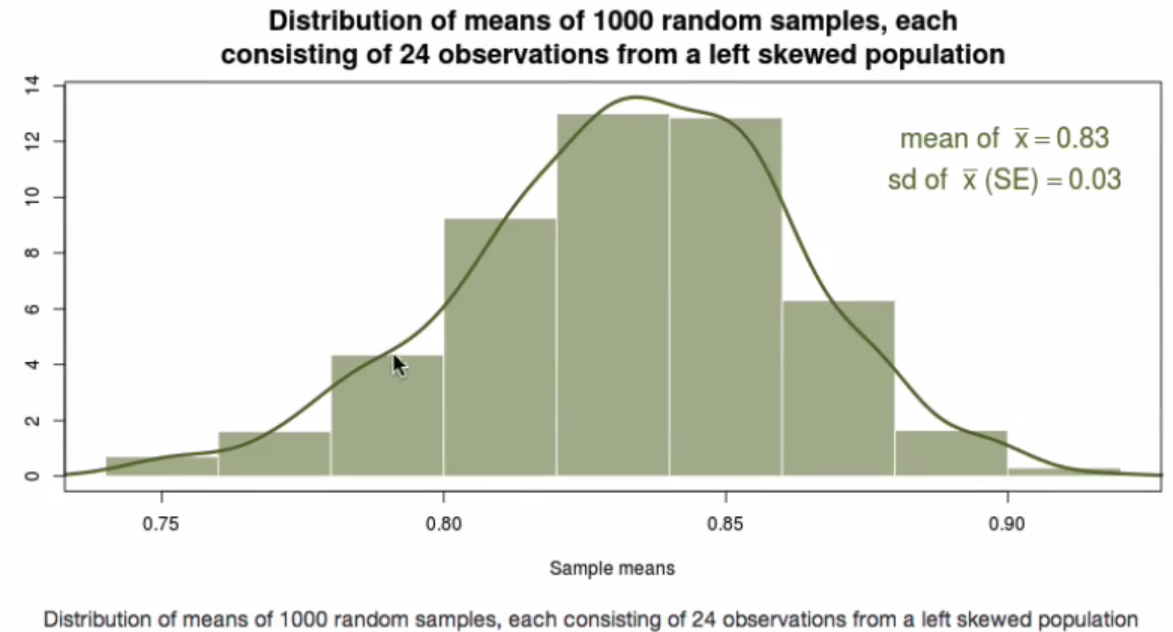




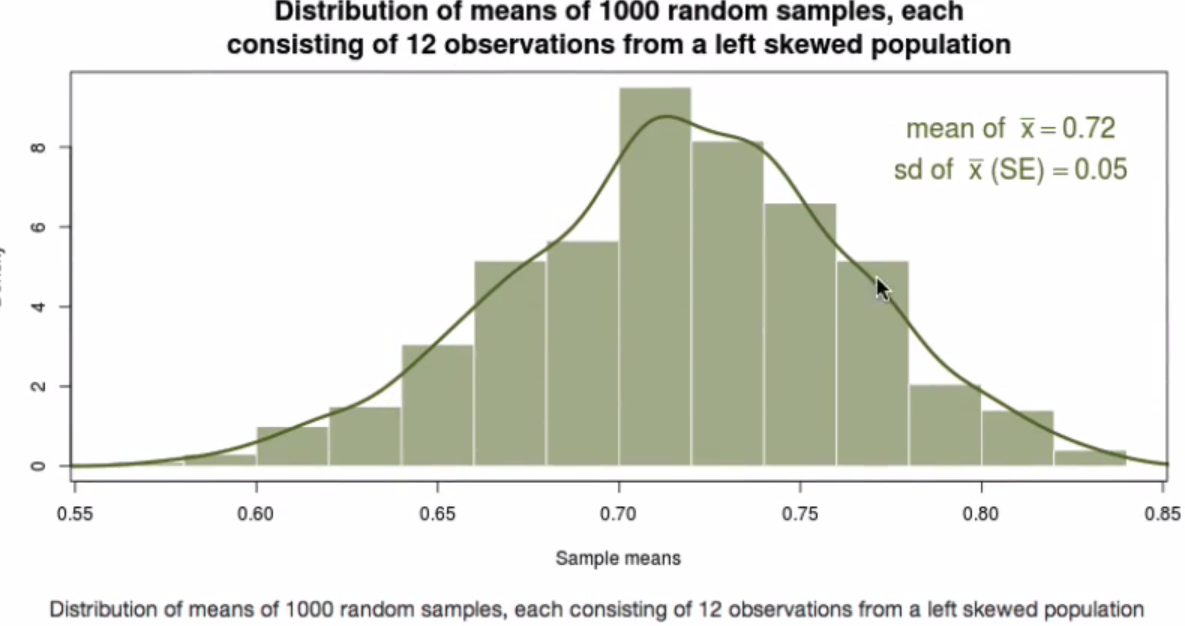
* 
* Sampling distribution = unimodal, symmetric, resembles a nearly normal distribution
* Population distribution that's left skewed w/ samples of size 500

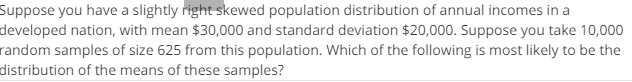
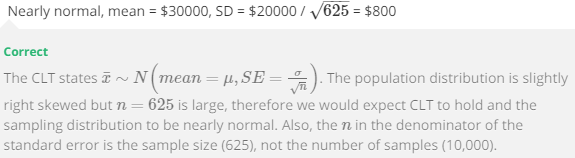


* Sampling distribution looks pretty nearly normal.
* Decrease sample size to 24



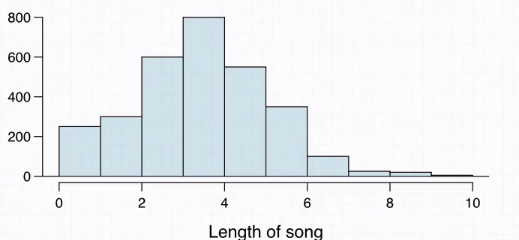
* Sampling distribution looks left skewed (like the population)
* Decrease n even more = even more skewed.
* If population distribution is not *that* skewed (low skew compared to high skew above) 🡪 don't need a whole lot of observations in a sample.



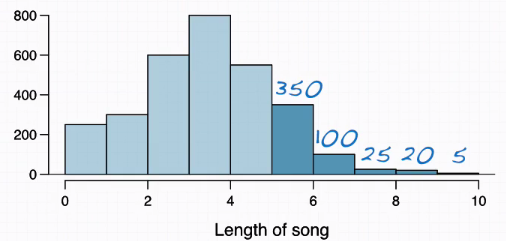
* Only 12 observations in each sample + sampling distribution = pretty unimodal + symmetric
* Moral of the story = **the more the skew, the higher the sample size needed for the CLT to kick in.**
* 
* 

**CLT (for means) examples**

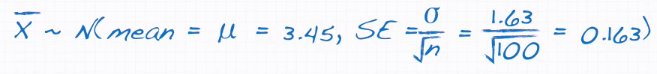
* Suppose my iPod has 3,000 songs 🡪 distribution of lengths of these songs.



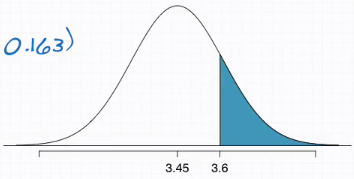
* **Mean length = 3.45 minutes + the SD = 1.63 minutes.**
* *Calculate the probability that a randomly selected song lasts more than 5 minutes*
* This is the same thing as saying “among the population of songs on this iPod, what % are > 5
* **Can only use Z-scores + the associated normal probabilities if the distribution is nearly normal**
* This is right-skewed 🡪 makes b/c a song can't be less than 0 minutes = natural boundary at lower end w/ really no upper end to how long songs can be.
* However, can assume it's going to be fewer counts of songs as the # of minutes increases.
* Can use the histogram + heights of the bars to estimate % of songs that fall between bins of minutes + use those to calculate probability.
* Interested in everything > 5 minutes 🡪 requires eyeballing heights of bars.



* Estimates might be slightly off, but should be w/in this range.
* Seems like there are no songs on this iPod last > 10 minutes.
* Let X = length of 1 song
* P(X > 5) = 350 + 100 + 25 + 20 + 5 / 3000 = 500 / 3000 = **0.17**
* So, the probability a randomly selected song on my iPod lasts more than 5 minutes = .17
* Another way of thinking about this is “17% of songs on my iPod last more than 5 minutes”.
* I'm about to take a trip to visit my parents + the drive is 6 hours + I make a random playlist of 100 songs. What is the probability my playlist lasts the entire drive?
* 6 hours = 360 minutes 🡪 need 360 minutes worth of songs.
* Need X1 + X2 + … + X100 >= 360 🡪 **P(X1 + X2 + … + X100 >= 360)**
* We haven't really worked w/ **sums of random variables**, but we know how to deal w/ averages.
* P(X1 + X2 + … + X100 >= 360) is equivalent to the average length of the 100 songs being > 360 divided by 100 = 3.6 minutes 🡺 **P(X\_ >= 3.6)**
* So we want the average length of the 100 songs to be greater than 3.6 minutes.
* This is NOT the same thing as *every single song* being more than 3.6 minutes
* Gives a *minimum* playlist length of 360 minutes.
* We just want the total to be greater than 360 minutes to last the entire drive.
* Now that we have **X­\_,** the **sample mean**, that should remind us that the CLT might be helpful
* Using the CLT, we can find the distribution of the sample mean pretty easily.
* The CLT says X\_ will be distributed nearly normally w/ mean = population mean, 3.45 min. + the standard error, SE\_x = the population SD, sigma, divided by the square root of n, the sample size



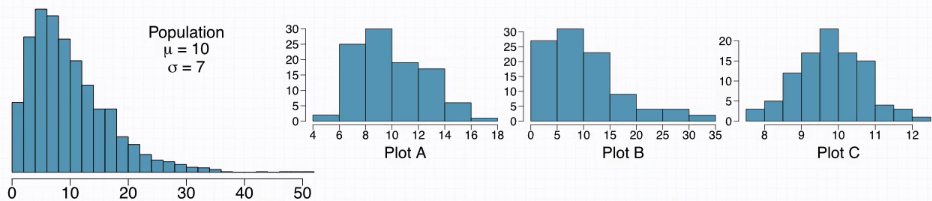
* Now, we:
* have a **random variable**, X\_, our sample mean
* know its distribution = normal
* know its mean = 3.45
* know something about its variability, the standard error (basically the SD of X\_) = 0.163
* are interested in *some* probability
* This combination of events = *a normal distribution, know its parameters, looking for probability,* should prompt that we 1st draw a curve before proceeding.

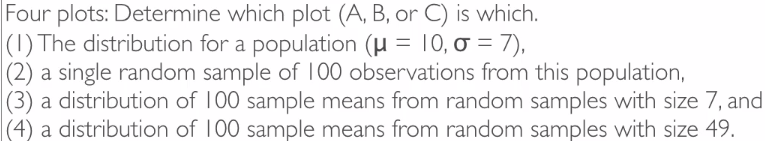


* Looking for the observation of interest as 3.6 minutes + everything above that.
* Remember, dr**awing the curve is always your friend.**
* Do this 1st = much less likely that you’d do something wrong in the following steps.
* Now calculate the Z-score 🡪 **X\_interest – X\_ / SE\_x** 🡪 **3.6 - 3.45 / .163** = **.92**.
* Note we divide by standard error + not sigma/SD of the population.
* This is b/c the observation of interest (3.6) *is a sample mean* + NOT an individual song/individual observation.
* **We measure the variability of INDIVIDUAL OBSERVATIONS w/ SDs.**
* **We measure the variability of SAMPLE MEANS w/ standard errors.**
* Whatever the observation is that you plug in in the numerator of a Z-score, *ITS OWN* variability belongs in the denominator.
* In other words, our observation = an X\_, + not an X.
* Can now easily find the area using many methods



* There's almost an 18% chance my playlist lasts *at least* the entire drive.
* 4 plots presented 🡪 Determine which plot is which of the following:



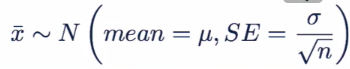


* 1st = 1st
* 2) = B 🡪 most like population = only 1 sample
* 3) = A
* 4) = C 🡪 greatest sample size = more normal distribution

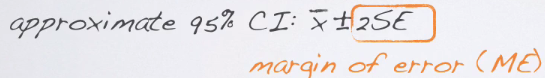
***4.1 CI’s***

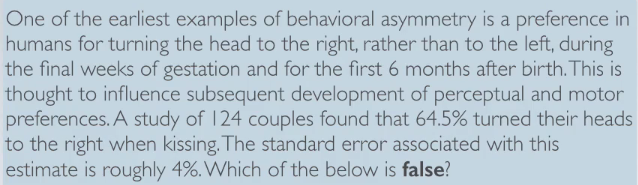
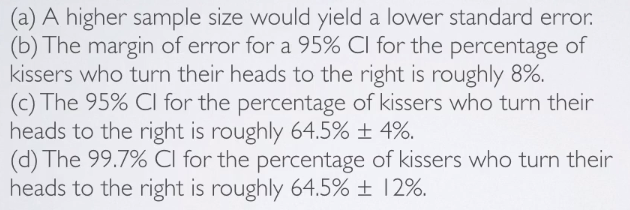
**CI (for a mean)**

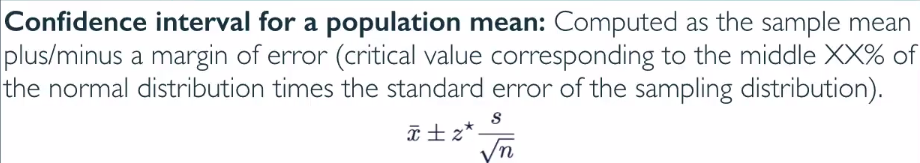
* **CI** = *plausible range of values for the population parameter*
* Using a sample statistic to estimate parameter = fishing w/ a spear
* Using a CI = fishing w/ a net
* Reporting **point estimates** = probably won’t hit parameter, but a range of possible values = better chance of encapsulating it
* Sample mean X\_ = best guess for parameter mu, so any CI is constructed around X\_
* Also, from CLT, we know X\_ is normally distributed w/ mean = mu + SE = sigma/sqrt(n)

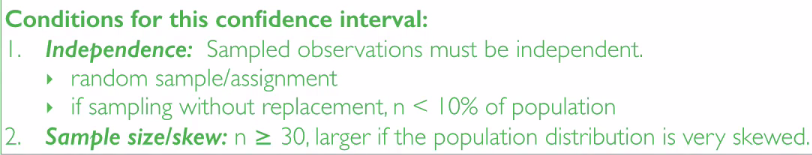
* For 95% of random samples, the unknown true population mean is w/in 2 SE’s of the sample mean
* 95% only applies to random samples in the abstract 🡪 once we get a sample, the mean of that sample will either be w/in 2 SE’s of the population mean or not
* CI’s are often a point estimate (X\_) +/- some margin of error (2\*SE for 95% CI)

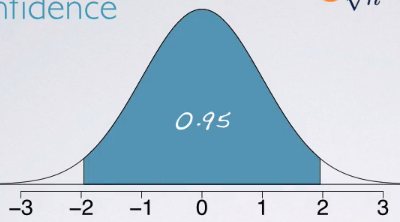


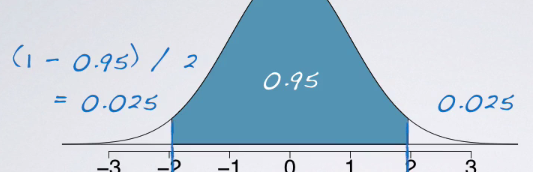
* 
* 
* A) is false b/c the higher the sample size, the less variable the point estimates from the samples



* *z\* should be approximately 2, according to 98/95/99 rule =* only rule of thumb
* The ability to use formula has same conditions as the CLT, since this method is based on the CLT



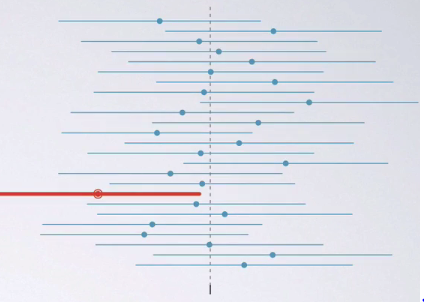
* Sample size for skew is stricter here w/ a *minimum* requirement for n 🡪 30
* So when checking our conditions, look at a visualization of the distribution from the sample we're going to use as an indicator for what the population looks like.
* Or be told we can assume some normality and proceed
* To find EXACT **critical value** of 95% CI, we know this encapsulate middle 95% of the data in the normal distribution
* 
* Want to find the **cutoff values** of the middle 95% 🡪 +/- 2.5%

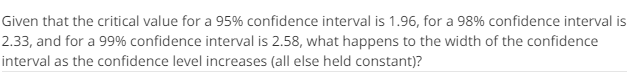




**Accuracy vs. Precision**

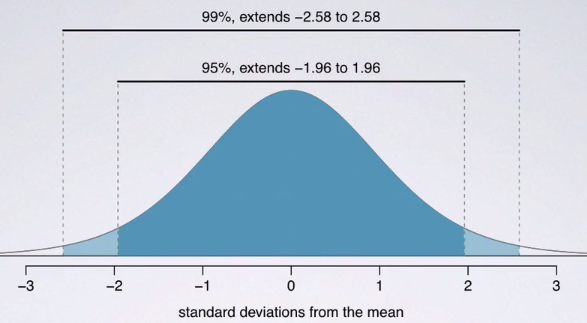
* **Accuracy =** whether or not the CI contains the true population parameter
* **Precision =** width of the CI
* Suppose we took many samples + built at CI from each using 
* 95% of those samples’ CI’s would be expected to contain the true population mean μ



* Vertical = μ each horizontal line = CI w/ point = point estimate at center
* Red = 1 sample whose CI did NOT contain μ 🡪 24/25 samples contain μ 🡪 **confidence level** for these intervals = .96 or 96%
* **Confidence level =** % of random samples that will yield CI’s that contain μ
* We usually only work w/ 1 sample from a population, so we CHOOSE the confidence level + base the rest of our calculations on it
* Common **confidence levels α =** .9, .95, .98, .99
* Adjusting **α** = adjusting critical value in the CI formula
* 

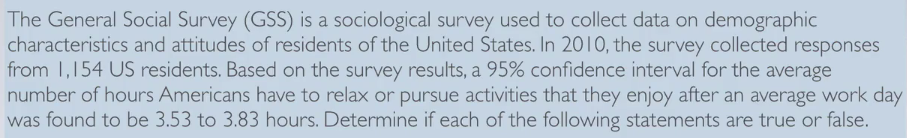
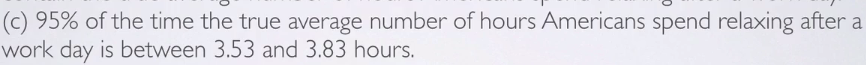
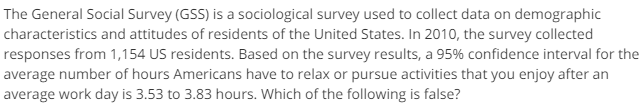
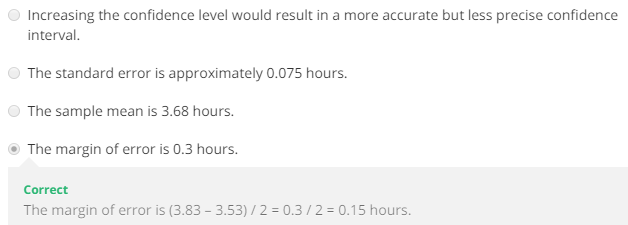


* To be very certain to capture **μ**, we want a wider interval/width of the AUC that captures the middle 95%/99% of the distribution



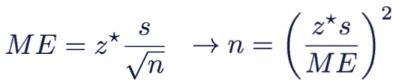
* More accurate = higher confidence level 🡺 may come at a **cost**
* W/ a wider interval, we are *less* ***precise***

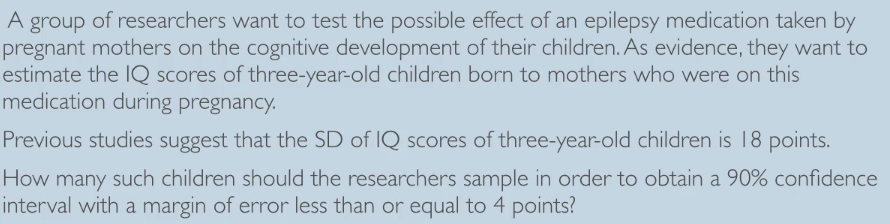


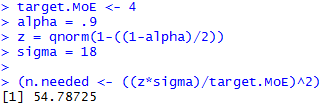
* So increase **α =** increases CI width = increases accuracy = decreases precision
* To get both better accuracy AND precision 🡪 increase n = shrinks SE and Margin of Error
* 
* 
* 
* 
* 
* \*\*\* **α** is about the POPULATION mean, NOT the SAMPLE mean
* Therefore, for D, could say 100% sure those in *that* *sample* spend that much time relaxing
* 
* 

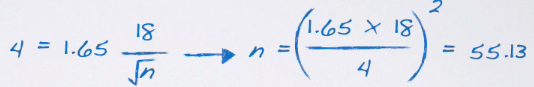
**Requires Samples Size for a Certain Margin of Error**

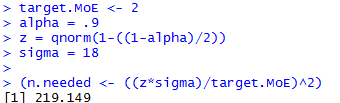
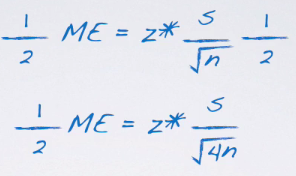
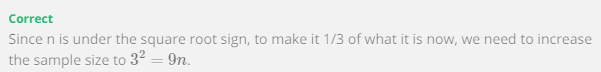
* Given target MoE, **α**, + info about the variability of a sample (or population), we can determine the required sample size n to achieve a desired MoE



* 





* *Minimum* required is 55, so we round up 🡪 need at least 56 children to achieve maximum margin of error of 4
* How would the required sample size change if we wanted MoE <= 2
*  
* *LONG STORY SHORT =* to decrease margin of error by 1/2, multiply sample size by 4
* 
* 

**CI (for the mean) Examples**