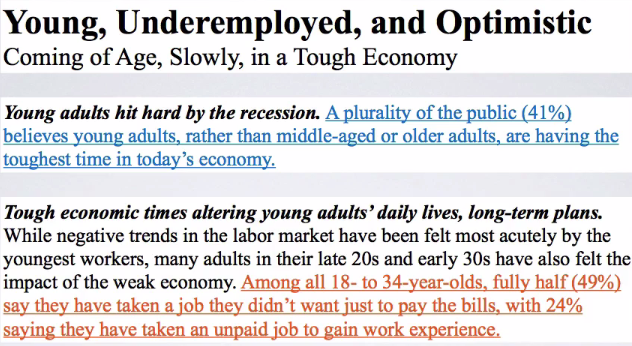
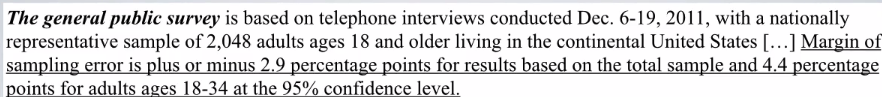
***COURSERA: STATS W/ R SPECIALIZATION***

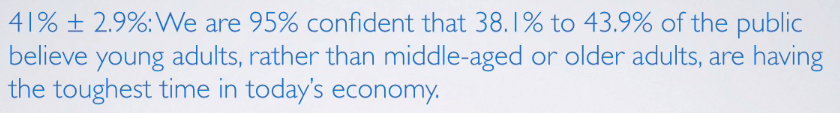
***COURSE 2 - Inference***

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

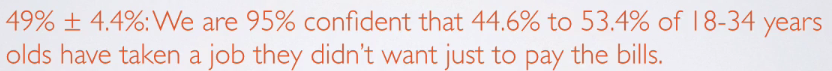
***4.1 CLT and Sampling***

**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.



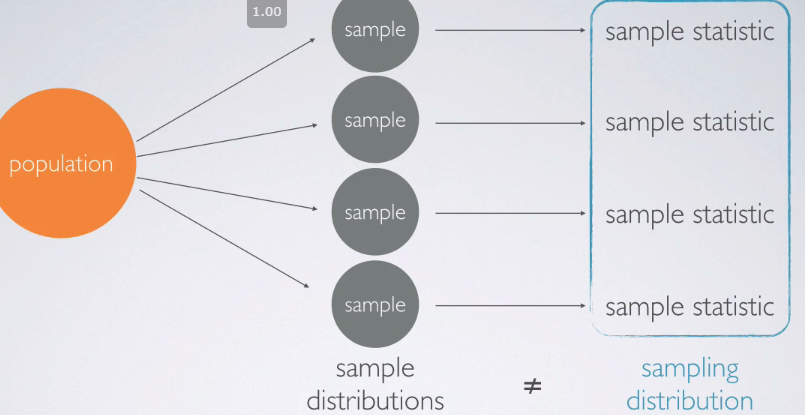
* We were also told that 49% of the public had taken a job they didn't want just to pay the bills.



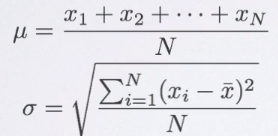
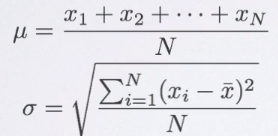
* The 41 + 49% we have on hand come from **sample statistics**, but we're often interested in **population parameters**.
* Since compete populations are difficult/impossible to collect data on (i.e. collecting data from the entire US population for this study), use sample statistics as **point estimates** 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



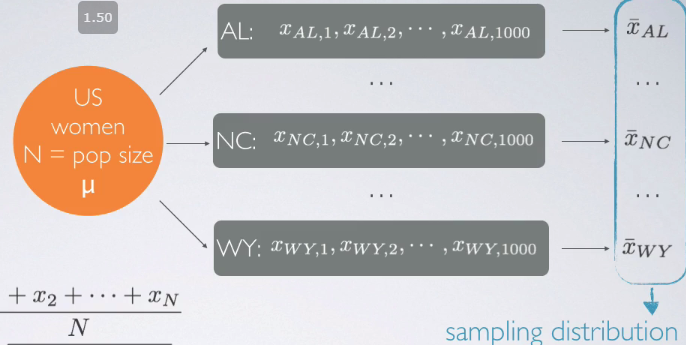
* Ex: Average height of US women 🡪 population of interest = US women w/ population size N, parameter = average height of ALL US women, μ
* We could calculate population mean + population SD **σ** (which wouldn’t be small since heights vary so much)

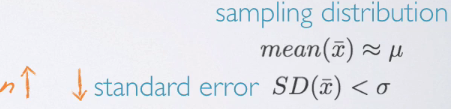
* Sample 1K women from each state, expecting the distribution of sample means from these samples to be less variable than the population
* Ex: Samples from Alabama where X is the observation and the # subscript is the i-th observation



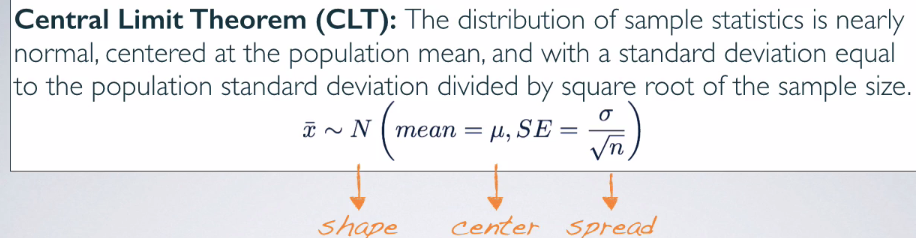
* Then find the mean for each state sample, X\_.
* All these state sample means = sampling distribution



* 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*
* 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
* 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.
* 1st, focus on the 10% condition 🡺 If sampling w/out replacement, n needs < 10% of the population
* Say you live in a very 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 something done often in surveys b/c we don't need your responses again), the probability of sampling you vs. somebody from your family would stay consistent throughout all trials + we wouldn't need to worry about the 10% condition there.
* 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.
* Say we have skewed population distribution here we have a population distribution that's extremely right skewed. When the sample size is small here we're looking at a sampling distribution created based on samples of n=10, the sample means will be quite variable. And the shape of their distribution will mimic the population. Distribution. Increase in the sample size a bit. Now we've gone from N equals 10 to N equals 100. This decreases the standard error, and the distribution starts to condense around the mean and starts looking more unimodal and symmetric.
* 16:17
* With quite large samples, here we're looking at our sampling distribution where for each of the individual samples based on which the sample means were calculated, those sample sizes were 200. With quite large samples like this, we can actually overcome the effect of the parent distribution. And the central limit theorem kicks in. And the sampling distribution starts to resemble a closely normal distribution. Why our we somewhat obsessed with having nearly normal sampling distributions? Because we've learned earlier that once you have a normal distribution, calculating probabilities which will later serve as our P values in our hypothesis tests are relatively simple. So, having a nearly normal sampling distribution that relies on central limit theory is actually going to open up a bunch of doors for us for doing statistical inference using confidence intervals and hypothesis test using normal distribution theory. Lets do another demo real quick. We looked earlier at what does a sampling distribution look like when we have nearly normal population distribution. Lets take a look to see what happens if the population distribution is not nearly normal.
* 17:30
* Suppose I first pick a uniform distribution. Here we can see, that our population distribution is uniform. Let's say that it is going to be uniform between four and obviously our upper bound needs to be greater, four and 12. So we can see a uniform distribution between four and 12, absolutely no peaks, so on and so forth. Say that we're actually taking samples of size just 15 from this distribution. Each one of our samples contains 15 observations from the parent population. And the center of these samples are going to be somewhere close to the population mean. We take a bunch of these samples, a thousand of them, and let's take a look at what the sampling distribution is looking like. It actually looks fairly symmetric. Unimodal and symmetric. The center of the distribution is very close to our population distribution mean. And the variability of this distribution is actually much lower than our population distribution. We can see that the standard error is .59. While the original population standard deviation was 2.31. What happens if we have skewed data? Here we have population distribution that's right skewed. We're taking samples of size 15, and let's actually make this an extremely right skewed distribution. So this is what this looks likes. If we're taking samples of size 15. So here, we're taking a look at each one of our individual samples. The sampling distribution is looking an awful lot skewed. However, if I increase my sample size, to be much larger, say 500, then my sampling distribution is starting to look much more unimodal and symmetric and starting to resemble a nearly normal distribution. What about a left skewed distribution? Once again, let's make the skew of this distribution pretty high, and we can see that our sampling distribution when we have a large number of observations in each sample. We have still kept it at 500 observations. The sampling distribution looks pretty nearly normal. However, if I was to decrease my sample size to be something pretty small, 24 let's say, then my sampling distribution is looking more and more left skewed. And in fact, if I take even smaller samples.
* 19:59
* Let's go all the way down to 12, for example. Now the distribution is looking even more skewed. If I though, decrease the skew and my population distribution to begin with is not looking all that skewed anyway, then I really don't need a whole lot of observations in my sample. Here I have only 12 observations in each sample. And the sampling distribution is already looking pretty unimodal and symmetric. So the moral of the story is, the more the skew, the higher the sample size you need for the central limit theorem to kick in. Please feel free to go play with this applet, interact with it, and find out for yourself what the sampling distribution looks like in various scenarios. And also play around with the different parameters of the distributions, either picking how skewed they are, if it's a uniform distribution, what the minimum and the maximum are? Or it's a normal distribution, what the mean and the standard deviation are?