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

***COURSE 2 - Inference***

**WEEK 4 - Inference for Proportions**

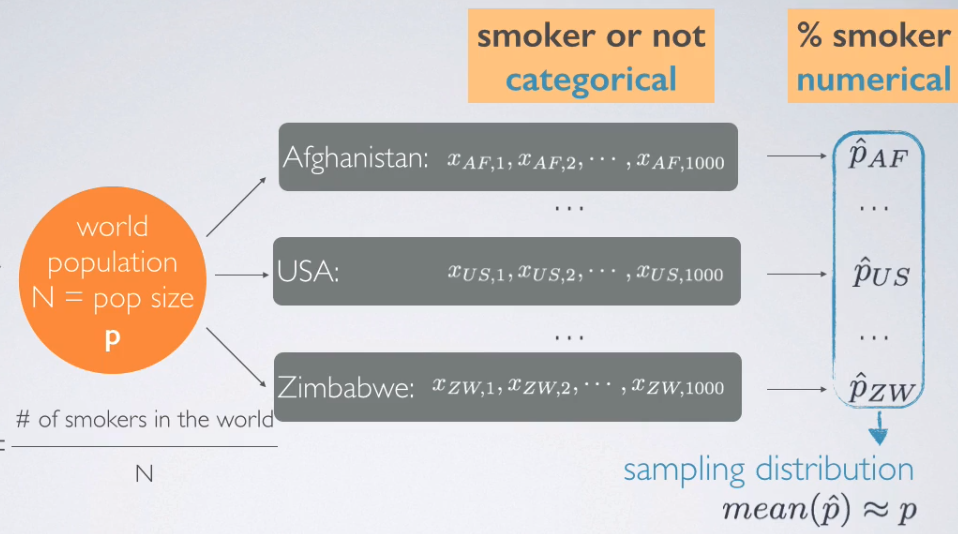
***4.4.1 Inference For Proportions***

**Introduction**

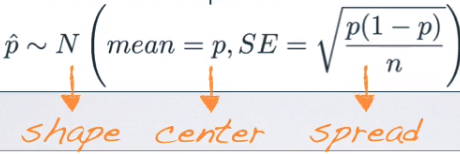
* Inference for categorical variables where **parameter of interest** = a proportion, as opposed to mean
* Gallup poll results where American public was asked about opinion on same-sex marriages
* Poll where data were collected from a variety of countries trying to answer the if most children in the country have the opportunity to learn + grow every day
* Published study on antihypertensive meds + serious falls on nearly 5k Americans > 70 during a 3 year period + found those who taking antihypertensive meds had a 30-40% greater likelihood of experiencing severe fall-related injuries like hip fractures + head trauma
* What is common between these studies = they deal w/ categorical variables like on same-sex couple marriages, whether children in a country have the opportunity to learn + grow every day, + whether patients taking a certain type of med are more likely to have fall-related injuries.
* Ex1: Simple case = categorical variable only has 2 levels we can categorize as a success or failure (**binary**)
* Success != something positive 🡪 could be a patient dying or suffering from a certain type of disease, or somebody graduating from high school
* Doesn't matter the context = important thing is these = **binary categorical variables** = levels can be categorized as either 1 thing or the other
* In this case, parameter of interest = **proportion of success**.
* Ex2: categorical variable w/ > 2 levels.
* Ex: Socioeconomic status tends to be categorized as low, medium, or high
* Ex3: 2 categorical variables that both have only 2 levels.
* IF somebody is male or female, if they decide to pursue a major in the sciences or not
* Ex4: 2 categorical variables, where either 1 or both of have > 2 levels
* Ex: Socioeconomic status (low, med, high) + educational attainment (finished high school, junior college, college, or graduate degree) + look at the relationship between these 2 variables
* In this case, we evaluate whether these variables appear to be dependent or independent.

**Sampling Variability and CLT for Proportions**

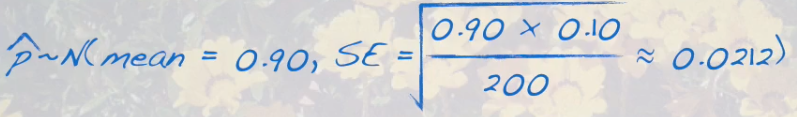
* **Sampling distribution for a sample *proportion***, not a mean, b/c we're dealing w/ categorical variables + **parameter of interest is no longer a mean but a proportion**.
* **CLT for proportions** = very similar to what before but w/ a different measure of the SE
* **Sampling Distribution:** Say you have a population of interest + take a random sample from it
* Based on that random sample, calculate a sample statistic + if variable of interest = categorical, the sample statistic = a sample proportion.
* Take another sample, calculate sample proportion from *that*, + then another, + so on
* Want to take as many samples as possible.
* Distributions of observations w/ in a sample = a **sample distribution**.
* However, when we look at the **distribution of the sample statistics** = **sampling distribution**
* Sample distributions = observations are individual (people, cases, etc.)
* Sampling distribution = observations = sampled statistics.
* Ex: Estimate proportion of smokers in the world, so population = world population, N = population size (everybody in the world), parameter of interest = **p**, the true proportion of smokers in the world
* If we actually had all population data, p = # of smokers in the world / total population size.
* Instead, take many samples from different countries (sample 1K people from Afghanistan + ask each if they’re a smoker or not + so on + so forth w/ many countries)
* Now we have a bunch of samples of 1k observations each, where observation = a person from that country + we summarize these samples.



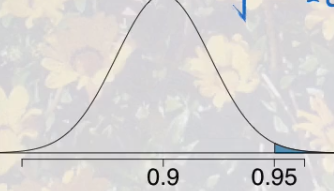
* Calculate proportion of smokers for each country (a sample proportion) + now the data set is not individual people but a data set of proportions
* The distribution of these proportions = the sampling distribution
* Each p^ = somewhat good guess for true p, although we expect variability between these b/c we’d expect trends in smoking habits of people from various countries.
* But overall, we’d expect the mean of the p^ values to be close to our unknown population mean, p
* So, started w/ a categorical variable (smoker or not) + for each sample, calculated a summary statistic = proportion of smokers + then create a distribution of numerical data = proportion of smokers in each country.
* started w/ a categorical variable + ended up w/ a distribution of a numerical variable
* **CLT says the distribution of sample proportions is going to be nearly normal** + just like w/ sample means, it's going to be **centered at the population proportion** (parameter) + w/ the **standard error inversely proportional to sample size** (also seen before)
* CLT tells us about the shape, the center, + the spread of a distribution



* SE = square root of p (**proportion of success**) times 1 – p divided by n.
* Conditions for the CLT for Proportions
* 1) **Independence of observations** = sampled observations must be independent + to achieve that either use random sampling or assignment, depending on study type
* + If sampling w/out replacement, make sure sample size < 10% of population.
* 2) **Sample size/Skew** = looking for a *balance of the sample size* *+ the proportion of success*.
* There should be at **least 10 successes + 10 failures in the sample**
* n\*p + n\*(1 - p ) must both be >= 10.
* Talked about this when we were dealing w/ the binomial distribution + were looking for the normal approximation of it
* Same idea holds here = want our sample proportion to be nearly normally distributed + therefore need to meet the **success failure condition** 1 more time.
* However, if p is unknown (for the both calculation of SE + for # of successes + failures), use our sample proportion, p^ b/c best guess for a population parameter = a sample statistic being used as a **point estimate** for that parameter.
* Ex: 90% of all plant species are classified as angiosperms (flowering plants). If you were to randomly sample 200 plants from the list of *all known* plant species, what is the probability at least 95% of the plants in sample will be flowering plants?
* Proportion of success = 0 + sample size, n, = 200
* Asked for probability of at least 95% successes 🡪 angiosperm plant = a success
* Looking for the probability our sample proportion will > 0.95.
* If we knew something about the distribution of p^, we should be able to easily calculate this + if we knew p^ was distributed nearly normally, we can calculate this probability using the normal distribution z scores + percentiles.
* CLT tells us it may be distributed nearly normally, so check to see if conditions for the CLT hold + if so, we can proceed w/ that.
* 1st Independence = 200 is certainly < 10% of all plants, so we can assume plants in our sample are independent of another.
* 2nd Success/Failure condition. n = 200 sample size + p = 0.9, so n\*p = 180 + n\*(1 – p) = 20
* Both of these are > 10, so success/failure condition holds
* **These 2 facts tell us the distribution of the sample proportion is going to be nearly normal** w/ mean = population parameter 0.90 + SE =



* We have a normal distribution, we know it's mean + it's variability (spread),+ we're looking for a probability associated w/ this distribution.
* 1st to do = draw our curve, mark the mean = 0.90, + then shade area of interest = anything beyond 0.95.



* To calculate this probability, we can refer to a z score.

> p.hat = .9 # mean of our sampling distribution = population parameter (when normal)

> n = 200

> (se <- sqrt((p.hat\*(1-p.hat))/n))

[1] 0.0212132

> (z <- (.95 - p.hat)/se)

[1] 2.357023

* We are more than 2 SDs away from the mean at this point, so it's going to be a pretty small probability that at least 95% of plants our sample of 200 will be flowering plants.

> # get proportion of values lower than this z-score on the curve

> pnorm(p.hat,.95,se)

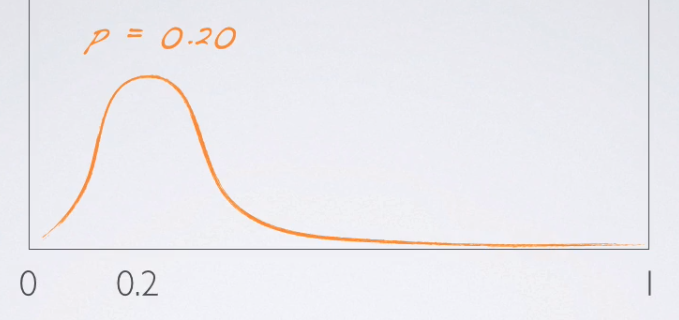
[1] 0.009211063

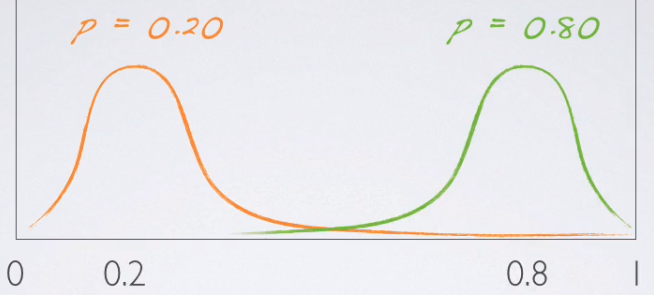


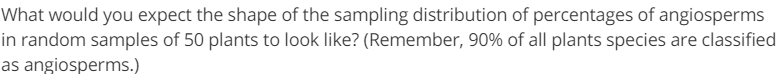
* We were looking for the probability of at least 95%, + so that seems like we should’ve used the notion p-hat >= 0.95.
* However, remember that **under a continuous distribution** (normal distribution is one), the probability of a random variable being equal to a specific number = 0
* So, we use the CLT to find this, but *could also do this using the binomial distribution as well.*
* Remember, sample size n = 200, proportion of overall success = 0.9, + we're being asked for p(obtaining 95% successes) = 95% of 200 = at least 190 successes in 200 trials where proportion of success is 0.9

> sum(dbinom(min.success:n,n,p)) # want prob of anything >= min of 195, up to 200

[1] 0.00807125

* That is not exactly the probability calculated before, but it’s awfully close
* **What if** the success failure condition is NOT met:
* Center of the sampling distribution will still be around the true population proportion + spread can still be approximated using the same formula for SE.
* However, the ***shape* of the distribution will depend on whether the true population proportion is closer to 0 or closer to 1**
* Remember distributions of proportions have natural boundaries = can only be between 0 + 1.
* So, we know that a sample proportion cannot <0 zero + or > 1
* Think about a situation where success/failure condition is not met + our true population proportion = 0.2, a value closer to 0 than to 1.
* We said the center of a sampling distribution is still around the true population parameter, but we end up w/ a smaller tail to the left of the distribution + a much longer tail to the right
* 
* This is b/c for samples taken from a population where true population proportion = 20%, we’d expect the majority of them to have sample proportions close to 20%, + we’ll still get some that are different than 20% (all the way down to 0 or all the way up to 1)
* But it's much less likely to get a sample proportion = 100% in a random sample from a population where true population proportion = 20%
* Left tail is short b/c we have a natural boundary at 0, but right tail is much longer b/c the natural boundary on the higher end doesn't appear until 1, so that yields a right-skew distribution.
* Similarly, if we had a true population proportion = 80%, we’d see the opposite effect



* This is only if the success/failure condition is NOT met.
* If the success/failure condition IS met, that means sample size is higher = will yield a smaller SE, so curves are going to be much more dense around the true population parameters + will look more + more symmetric as the sample size increases.
* 
* 