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

***COURSE 2 - Inference***

**WEEK 4 - Inference for Proportions**

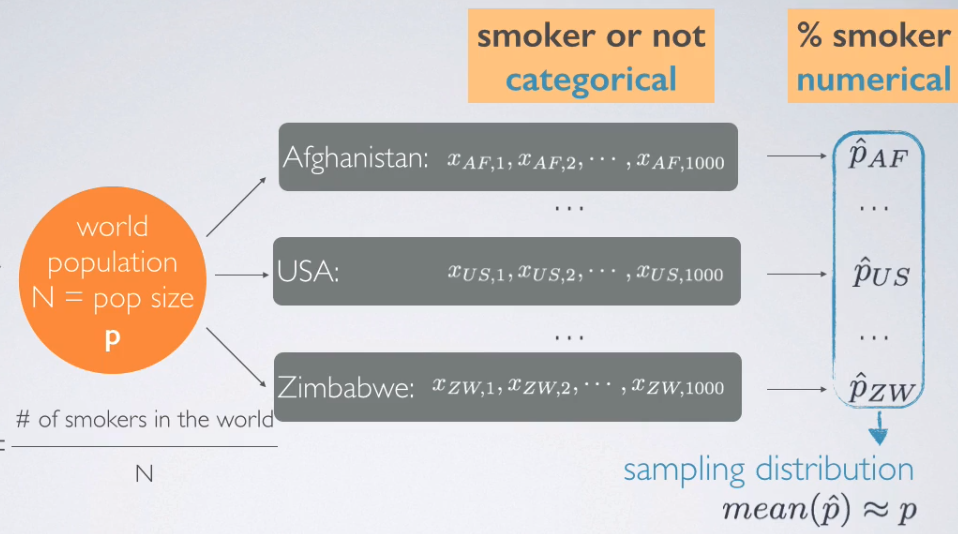
***4.4.1 Inference For Proportions***

**Introduction**

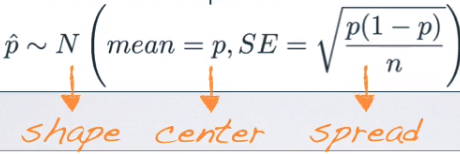
* Inference for categorical variables where **parameter of interest** = a proportion, as opposed to mean
* Gallup poll results where American public was asked about opinion on same-sex marriages
* Poll where data were collected from a variety of countries trying to answer the if most children in the 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 + CLT for Proportions**

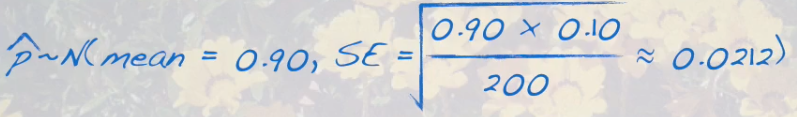
* **Sampling distribution for a sample *proportion***, not a mean, b/c we're dealing w/ categorical variables + **parameter of interest is no longer a mean but a proportion**.
* **CLT for proportions** = very similar to what before but w/ a different measure of the SE
* **Sampling Distribution:** Say 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.



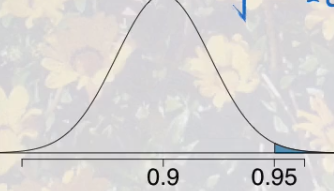
* Calculate proportion of smokers for each country (a sample proportion) + now the data set is not individual people but a data set of proportions
* The distribution of these proportions = the sampling distribution
* Each p^ = somewhat good guess for true p, although we expect variability between these 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)

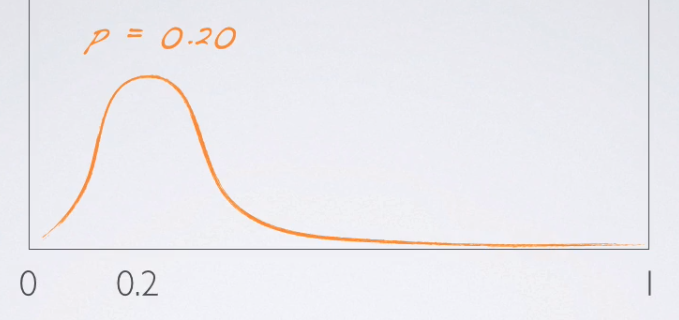
[1] 0.009211063

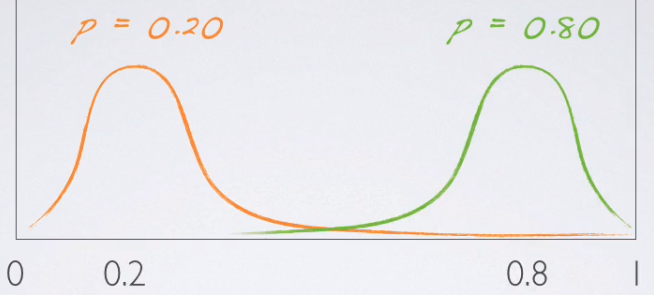


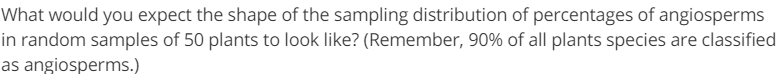
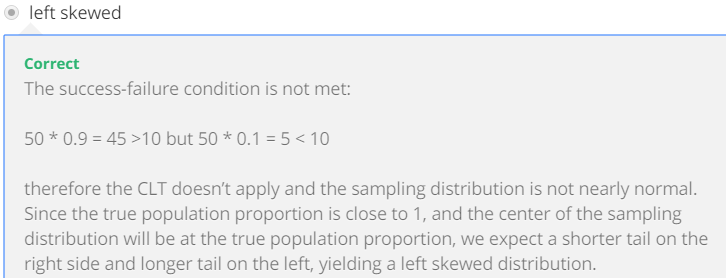
* We were looking for the probability of at least 95%, + so that seems like we should’ve used the notion p-hat >= 0.95.
* However, remember that **under a continuous distribution** (normal distribution is one), the probability of a random variable being equal to a specific number = 0
* So, we use the 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.
* 
* 

**CI for a Proportion**

* 2 scientists want to know, if a certain drug is effective against high BP. The 1st scientist wants to give the drug to 1K people w/ high BP + see how many experience lower BP levels.
* The 2nd scientist wants to give the drug to 500 people w/ high BP + NOT give the drug to another 500 people w/ high BP + see how many in both groups experience lower BP levels.
* *Which is the better way to test this drug?*
* We know **controlling** is important when running experiments, the 2nd study where the group that doesn't get the drug acts = the **control group**, should be the better design.
* This question was posed to 670 Americans w/in the GSS in 2010, + 99 said all 1K should get the drug.
* So, we're going to be categorizing these as those w/ a “bad intuition” for experimental design
* 571 said 500 should + shouldn’t get drug 🡪 label “good intuition” about experimental design
* Our goal is to **estimate what % of Americans have good intuition about experimental design**.
* **Parameter of interest** = % of all Americans who have good intuition about experimental design, + denote this unknown population parameter, p, for population proportion.
* Our **point estimate** = % of SAMPLED Americans who have good intuition about experimental design, denoted p^, our KNOWN **sample proportion**

> n <- 670

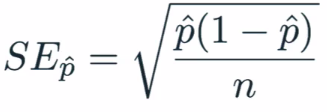
> bad <- 99

> good <- 571

> (p.hat <- good / n) # point estimate

[1] 0.8522388

* When it comes to estimation of an unknown population parameter, it always follows the same structure 🡺 **point estimate +/- a margin of error**.
* In this case our point estimate = sample proportion, p^ + margin of error = z\* (critical value) times SE of p-^
* So, once again, *the only new concept here = how to calculate the SE for the sample proportion*
* To calculate SE for PROPORTIONS, use the formula based on the CLT for PROPORTIONS:



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

[1] 0.01370956

* We don't know p (true population parameter) so we plug in the sample proportion, p^
* In most instances, we truly don't know the true population parameter + that’s why we're calculating a CI in the first place.
* So, roughly 85% of Americans, answered the question on experiment design correctly + we are asked to estimate using a 95% CI for the proportion of all Americans who have good intuition about experiment design
* **Before we calculate the CI, make sure that conditions for inference have been met**
* 1: Independence: Relies on a random sample/assignment + < 10% of the population being sampled.
* 670 Americans is definitely < 10% of all Americans + we know the GSS samples randomly.
* Therefore, we can assume that whether an American in the sample has good intuition about experimental design is independent of another.
* 2: Sample Size/Skew: check this condition when dealing w/ *categorical* variables + *proportions* as the success/failure conditions.
* Need to make sure that we have at least 10 successes + 10 failures in our sample.
* The sample size overall is large so we should be good here but let's take a look real quick.
* successes <- good # 571
* > failures <- bad # 99
* > (successes >= 10 & failures >= 10)
* [1] TRUE
* We didn't have to even go through the n\*p^ route here b/c we already know the # of successes + failures + know both of these #’s are indeed > 10.
* Therefore, since the success-failure condition is met, we can assume the sampling distribution of the proportion is nearly normal.
* Now that we have all of our building blocks we can actually calculate our CI.

alpha = .95

> (z.star = qnorm(1-((1-alpha)/2)))

[1] 1.959964

> (mOe <- z.star\*se.prop)

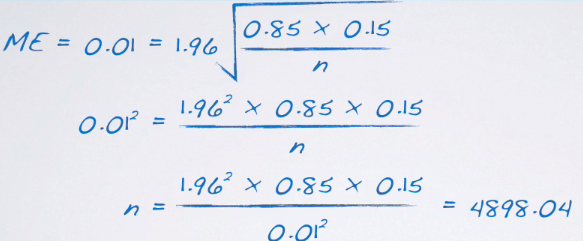
[1] 0.02687024

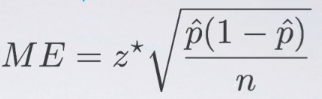
> (lower <- p.hat - mOe)

[1] 0.8253686

> (upper <- p.hat + mOe)

[1] 0.879109

* We can interpret this as, **we are 95% confident that 82.3% to 87.7% of all Americans have good intuition about experimental design**.
* Also, **95% of random samples of 670 Americans will yield CIs that contain the true proportion of Americans who have good intuition about experimental design**
* The margin of error for this CI was 2.7%, + if we wanted to reduce margin of error to 1% while keeping confidence level the same, at least how many respondents should we sample?
* > desired.mOe <- .01
* > (min.n <- ((z.star/desired.mOe)^2)\*(p.hat\*(1-p.hat)))
* [1] 4837.465
* 
* Remember we need to round this up even though mathematically, it doesn't make sense, b/c this is saying is that “in order to ensure a maximum 1% margin of error, we need 4,898.**04** persons.
* We can't have 0.04 of a person, so we’d need at least 4,899 people in our sample.
* So, for a minor reduction in our margin of error, we have to increase our sample size a lot, b/c the sample size appears *under the square root sign in calculation of the margin of error*.
* So, to have benefits from an increased sample size, increase your sample size by a lot before you can actually start reaping the benefits.
* 1 more point about calculating required sample size for a desired margin of error.
* Remember the formula for the margin of error = a critical value times the standard error.



* If there is a previous study we can rely on for the value of p^ in this formula, we’d use that in calculation of the required sample size (what we just did)
* *If not*, use 0.5 for p^ 🡺 2 reasons why we do this
* 1) if you don't know any better + it’s a categorical variable w/ 2 outcomes (S/F), 50-50 = a pretty good guess.
* 2) Using 0.5 for p^ gives the most conservative estimate = the highest possible sample size.
* We like being conservative when it comes to estimating minimum required sample sizes, b/c we definitely don't want to make a mistake + have to re due our sampling.

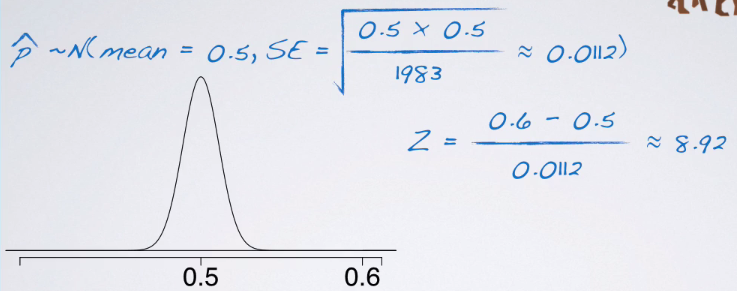
**Hypothesis Test for a Proportion**

* 1) Set our hypothesis:
* H0 = unknown population parameter p (opposed to mu for means) is set to some null value, +
* H1 = p can be <, >, or != to that null value.
* 2) Calculate point estimate = sample proportion, p^
* 3) Check conditions
* 1) Independence = make sure sampled observations are independent of each other (ensured either through random sampling or random assignment Depending on study type
* A) if sampling w/ a replacement, want sample size to be < 10% of the population.
* 2) Sample Size/Skew = want to make sure we have at least 10 expected successes + expected failures in our sample.
* NOTE: We used p, instead of p^ for this b/c **in a hypothesis test, we have to assume the null is true.**
* Think about definition of a p value = probability of observed or more extreme outcome if the null was true.
* So, when going through the conditions (or any other portion of the hypothesis test), we MUST assume the null is true, + therefore, wherever we see a p, plug in the null value for p
* So, we read this as not 10 *observed* successes + failures, but instead as 10 *expected* successes + failures.
* 4) Draw the sampling distribution before calculating p value + shade where the p-value belongs to.
* Either in one tail (upper or lower) or in a 2 tails
* 5) Calculate test statistic = *ALWAYS* of the form (observed – null) / standard error.
* This is observed sample proportion **p^** - null value **p** comes from null divided by **SE (**square root of p\*(1 – p) / n.
* Note again we said **p** + NOT p^, b/c we’re **assuming the null is true + therefore are using what the null has set forth as our true population parameter.**
* We don't *know* if that's the case, but we *must assume* the null is true as we proceed through a hypothesis test.
* 6) Make a decision + interpret it in context of the research question.
* If p < our significance level, reject the null + decide that the data provide convincing evidence for the alternative.
* If, in fact p > our significance level, we fail to reject the null + conclude the data do not provide convincing evidence for the alternative.
* Moral of the Story = use Sample proportion when there's nothing else known + use the population proportion (or at least the null-hypothesized value of the population proportion) when doing a hypothesis test as they dictate that we must assume the null is true.
* If checking the success-failure condition for a **CI**, use *observed sample proportions*.
* If calculating SE for a CI, use observed sample proportion, b/c we don't know any better
* If, checking the success-failure condition for a **hypothesis test** = use *expected counts* + plug in the p that comes from the null.
* If I'm calculating SE for a hypothesis test, SE = square root of p\*(1 – p) / n where p comes from the null.
* Ex: 2013 Pew Research poll found 60% of 1,983 randomly sampled American adults believe in evolution. Does this provide convincing evidence that the majority of Americans believe in evolution, where majority is > 50%.
* So if the question is “Is the *true proportion* of Americans who believe in evolution > 50%”, then our alternative H1 is “p > 0.5”
* Using this, we can easily figure out what the null can be, b/c we keep the same population parameter (.5) + same null value, except we simply set it equal to that # as opposed to giving a direction (1 way or another) or saying !=
* Remember, the null always has an equal sign in it vs. the alternative could have >, <, or !=, depending on the research question being posed.
* We are also given a sample proportion p^ = 0.6.
* So, *in this sample*, > 50% of respondents believe in evolution, but we're checking to see if this observed difference between the sample proportion + what we're hypothesizing is statistically significant.
* In other words, does *this particular sample* yield convincing evidence of majority of Americans believing in evolution?
* Before we move on to actually doing inference, always check conditions.
* 1) Independence: n = 1,983 is definitely < 10% of all Americans + we have a random sample, therefore we can assume an American is independent of another.
* 2) Sample Size/Skew of sampling distribution.
* **For proportions, check this using the success-failure condition**
* B/c we're doing a hypothesis test, we have to assume the null is true, so we’d use the p set forth by the null
* So, the total # of expected successes + failures in this sample both = 1983\*0.5 = 983 >= 10
* Since both conditions are met, we can assume a nearly normal sampling distribution for our sample proportion +, given a set of hypotheses + characteristics of sample, we can calculate our p-value.
* But before that, we need a test statistic + before *that*, we need to draw the sampling distribution.
* Our p^ is distributed nearly normally according to our conditions + to this CLT for proportions
* The center of that distribution should be @ the true population parameter, which we don't know
* But, since we’re doing a hypothesis test, we are assuming the null is true, so we plug in the value of the population parameter set forth in the null
* **Assume this is indeed the true population parameter for the purpose of this hypothesis test.**

(se.prop <- sqrt((h0\*(1-h0))/n))

[1] 0.01122816

* Next, we draw our sampling distribution centered at the null value + shade everything beyond 0.6 ( observed sample proportion, p^



* Then calculate our test statistic

> (z <- (p.hat - h0)/se.prop) # observed - null / n

[1] 8.906178

* That's a pretty high test statistic, if you think about it, b/c it's much farther than 3 SDs from the mean
* So the p value (AUC area under the z curve beyond 8.92) is going to be almost zero.

pnorm(h0,p.hat,se.prop)

[1] 2.641113e-19

* W/ such a small p value, the conclusion is going to be to reject the null