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

***COURSE 2 - Inference***

**WEEK 4 - Inference for Proportions**

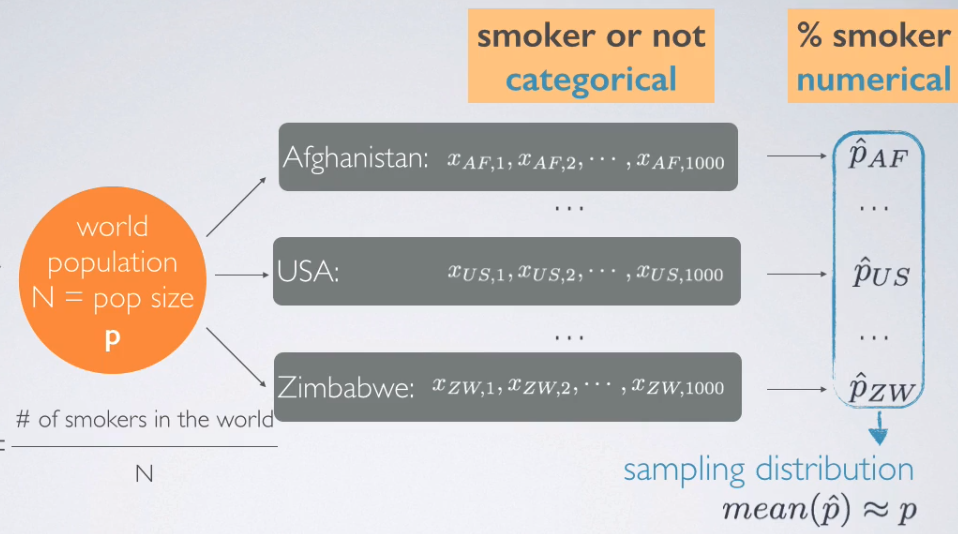
***4.4.1 Inference For Proportions***

**Introduction**

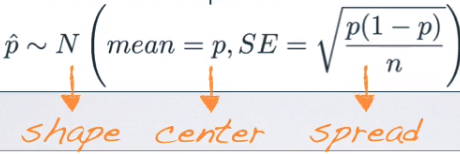
* Inference for categorical variables where **parameter of interest** = a proportion, as opposed to mean
* Gallup poll results where American public was asked about opinion on same-sex marriages
* Poll where data were collected from a variety of countries trying to answer the if most children in the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

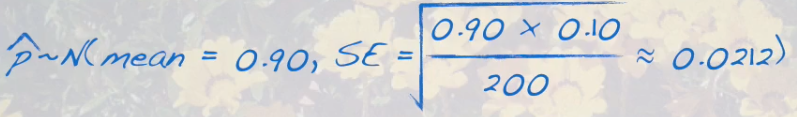
* **Sampling distribution for a sample *proportion***, not a mean, b/c we're dealing w/ categorical variables + **parameter of interest is no longer a mean but a proportion**.
* **CLT for proportions** = very similar to what before but w/ a different measure of the SE
* **Sampling Distribution:** Say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



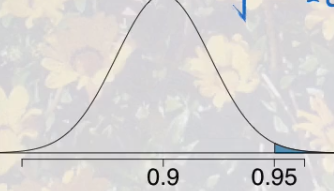
* Calculate proportion of smokers for each country (a sample proportion) + now the data set is not individual people but a data set of proportions
* The distribution of these proportions = the sampling distribution
* Each p^ = somewhat good guess for true p, although we expect variability between these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

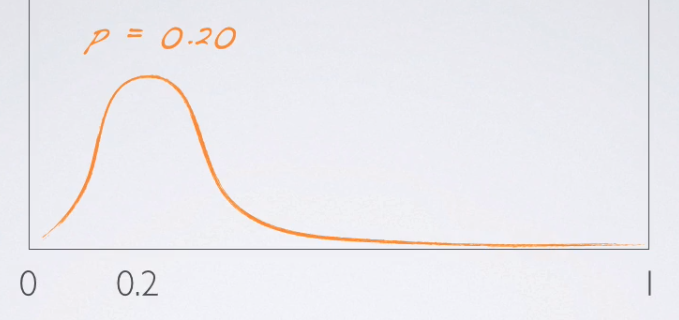
[1] 0.009211063

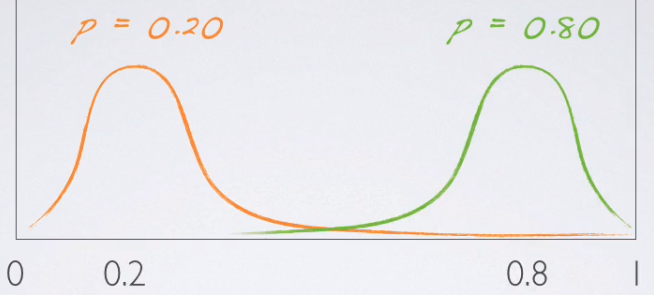


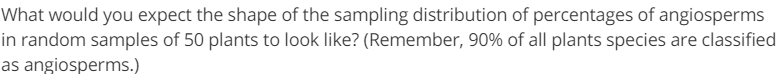
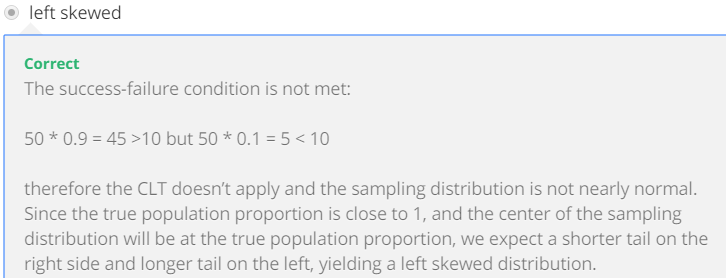
* We were looking for the probability of at least 95%, + so that seems like we should’ve used the notion p-hat >= 0.95.
* However, remember that **under a continuous distribution** (normal distribution is one), the probability of a random variable being equal to a specific number = 0
* So, we use the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

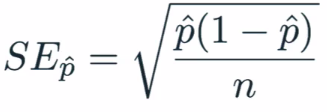
> bad <- 99

> 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)

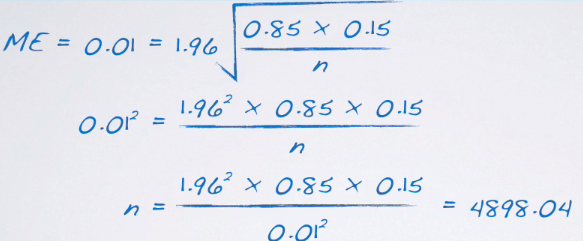
[1] 0.02687024

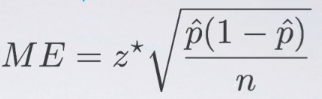
> (lower <- p.hat - mOe)

[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.

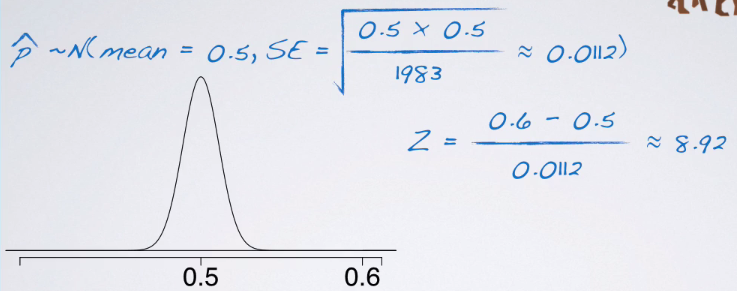
**Hypothesis Test for a Proportion**

* 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.

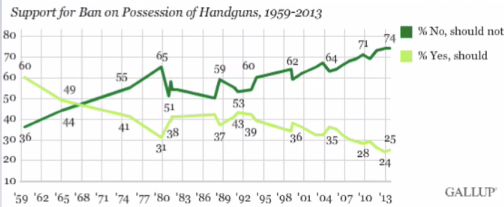
pnorm(h0,p.hat,se.prop)

[1] 2.641113e-19

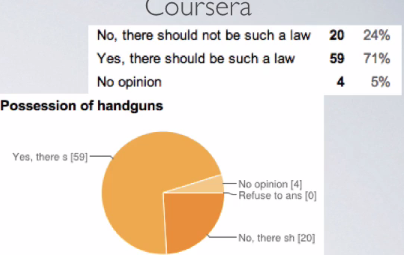
* W/ such a small p value, the conclusion is going to be to reject the null

**Estimating the Difference Between Two Proportions**

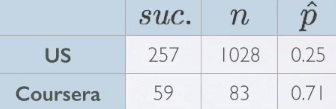
* Now we’re working w/ 2 categorical variables + evaluating the relationship between them.
* 1 categorical variable = explanatory/grouping variable + the other = response variable.
* Both have 2 levels, so a success + a failure 🡪 calculate proportion of successes in the 2 groups, based on our sample, + compare them to be able to say something about the proportions of their successes in the population
* This an *estimation* = use a CI for the difference between the 2 population proportions (unknown) using data from our samples.
* Early October 2013, Gallup poll asked, "Do you think there should/should not be a law that would ban the possession of handguns, except by the police + other authorized persons?"
* Possible responses = **no**, shouldn’t be such a law, **yes**, should be such a law, or **no opinion**
* Categorize **yes = success**, + either **no opinion or no = failure**.
* Gallop has been asking this since 1959 🡪Look to see how things have changed in the U.S. since then



* Back in the 50s or early 60s = higher proportion of people thinking there should be such a law, + but a much lower proportion today at 25%, compared to 74% that say there shouldn’t be such a law
* 74 + 25 = 99% so that 1% = no opinion on the matter.
* Also asked this question to a group of Coursera students



* This seems like a very different population 🡪 probably expected b/c Coursera is a much more international population than the US + possession of guns tends to be a hot topic in the US + may not necessarily be so in other countries.



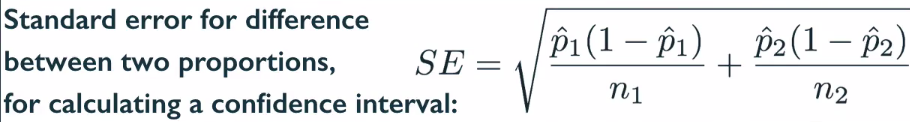
* Oct 2013 = Gallup surveyed 1,028 people + 25% said yes = 257 people
* In the Coursera population, 83 sampled + 71% said yes = 59 respondents
* *How do Coursera students + the American public at large compare w/ respect to their views on laws banning possession of handguns?*
* Parameter of interest = ***difference* between proportions of ALL Coursera students + ALL Americans who believe there SHOULD be a ban on possession of handguns.**

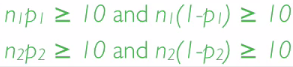
* p = unknown population parameter = population proportion
* p^ = sample statistic = serves as **point estimate** for parameter of interest = difference between the proportions of *sampled* Coursera students + *sampled* Americans who believe in a ban
* To estimate differences between proportions 🡪 estimation is usually a CI + a **CI** is always of the form **point estimate +/- a margin of error.**
* Point estimate = difference between 2 sample proportions
* Margin of error = critical score (Z\* b/c dealing w/ sample proportions ,+ based on the CLT, we know the distribution of the difference between the sample proportions are going to be nearly normal) times SE of the difference between the 2 sample proportions.



* To calculate the SE for the difference between 2 proportions for calculating a CI



* This is different for CI’s vs. hypothesis tests.
* Bringing 2 unknowns together (sample proportions from 2 groups), so overall variability should increase + hence we're adding the 2 variability components here
* Then we the square root of that to go from variance to SD, or, the SE (SD of the sampling distribution)
* Conditions
* 1) Independence.
* W/in groups 🡺 make sure sampled observations are independent of each other
* Ensured w/ random sample or assignment, depending on whether study type
* If we are sampling w/out replacement, sample size must be < 10% of population.
* Between groups 🡺 make sure groups are independent of each other (data are not **paired**)
* 2) Sample Size/Skew 🡺 ensure normality of the sampling distribution
* EACH sample should meet the **success-failure condition**.



* Remember for a CI, proportions considered to calculate success-failure condition = *observed* sample proportions.
* Using a 95% CI, estimate how Coursera students + the American public at large compare w/ respect to their views on laws banning possession of handguns
* Independence condition 🡪 DO have a random sample for US population, but *not for Coursera population* =simply a voluntary poll posted on Coursera discussion forums, so can't say it was a random sample.
* 10% condition was met for both of the groups = both 1028 + 83 are < 10% of their respective populations
* This means sampled Americans can be assumed to be independent of each other but sampled Coursera students *may not be*.
* So in this case, be wary of generalizing any of conclusion from these findings to the overall population at large b/c we don't really have a good sample from the Coursera population.
* Still move on w/ our analysis only for illustrative purposes
* Sample size + skew condition. 🡪 can we assume the sampling distribution of the *difference* between the 2 proportions is nearly normal
* When dealing w/ proportions, check for this using **success failure rule** + when doing a CI, check for this w/ **observed** successes + **observed** failures
* US = 257 observed successes + (1,028 – 257) = 771 observed failures
* Coursera = 59 successes + 24 failures
* Both of these met = success/failure condition is met for both of the groups + therefore we can assume the sampling distribution of the difference between the 2 proportions is nearly normal
* Now that we got our conditions out of the way, we can actually calculate our CI.

(point.estimate <- p.hat.c - p.hat.a)

[1] 0.4608434

> (point.estimate <- p.hat.c - p.hat.a)

[1] 0.4608434

> alpha = .95

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

[1] 1.959964

> (se.diff <- sqrt((p.hat.c\*(1-p.hat.c))/n.c + ((p.hat.a\*(1-p.hat.a))/n.a)))

[1] 0.05156394

> (point.estimate <- p.hat.c - p.hat.a)

[1] 0.4608434

> alpha = .95

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

[1] 1.959964

> (se.diff <- sqrt((p.hat.c\*(1-p.hat.c))/n.c + ((p.hat.a\*(1-p.hat.a))/n.a)))

[1] 0.05156394

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

[1] 0.1010635

> (lower <- point.estimate - mOe)

[1] 0.3597799

> (upper <- point.estimate + mOe)

[1] 0.5619068

* We are 95% confident the proportion of Coursera students who believe there should be a ban on possession of handguns is 36-56% higher than the proportion of Americans who do.
* That's a huge difference, even when we factor in the variability around the point estimate
* Again, this is probably expected based on how different the composition of the 2 populations are.
* **Does the order of proportions in our calculations matter**? = Yes + No.
* Might change some calculations along the way, but not THE conclusions
* Margin of error is bound to be always positive b/c SE is always going to be positive.
* Conceptually speaking, “negative variability” doesn't make sense + mathematically speaking, SE = a quantity calculated as a square root
* Z\*, by definition, is always going to be positive as well.
* On the other hand, point estimate can be positive/negative, depending on the order of subtraction

point.estimate <- p.hat.a - p.hat.c)

[1] -0.4608434

> alpha = .95

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

[1] 1.959964

> (se.diff <- sqrt((p.hat.a\*(1-p.hat.a))/n.a + ((p.hat.c\*(1-p.hat.c))/n.c)))

[1] 0.05156394

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

[1] 0.1010635

> (lower <- point.estimate - mOe)

[1] -0.5619068

> (upper <- point.estimate + mOe)

[1] -0.3597799

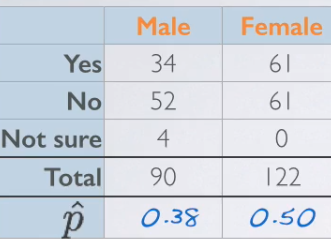
* These are actually exactly the same as the 1st interval 🡪 We are 95% confident the proportion of Americans who believe there should be a ban on possession of handguns is 36%-56% lower than the proportions of Coursera students
* As long as your interpreting correctly, it doesn't matter in which order you do the calculations.
* 1 last step 🡺 Based on the CI calculated, *should we expect to find a significant difference between the population proportions of Coursera students + the American public at large who believe there should be a law banning the possession of handguns at the equivalent significance level?*
* In this case, we’re asked to do a kind of mock hypothesis test, + in that case, our null = the difference between the 2 population proportions = 0 = these 2 populations are exactly the same
* For our CI, anything between .36 + .56 is fair game for the difference between the 2 population proportions
* The value= 0 does not appear, + based on that, we’d reject this null + say that “based on this CI, it doesn't appear that these 2 populations are the same w/ respect to proportion of those who believe there should be a law banning the possession of handguns”

pnorm(point.estimate,.95,se.diff)

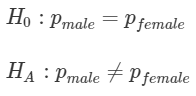
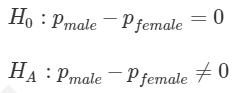
[1] 3.989529e-165

**Hypothesis Test for Comparing Two Proportions**

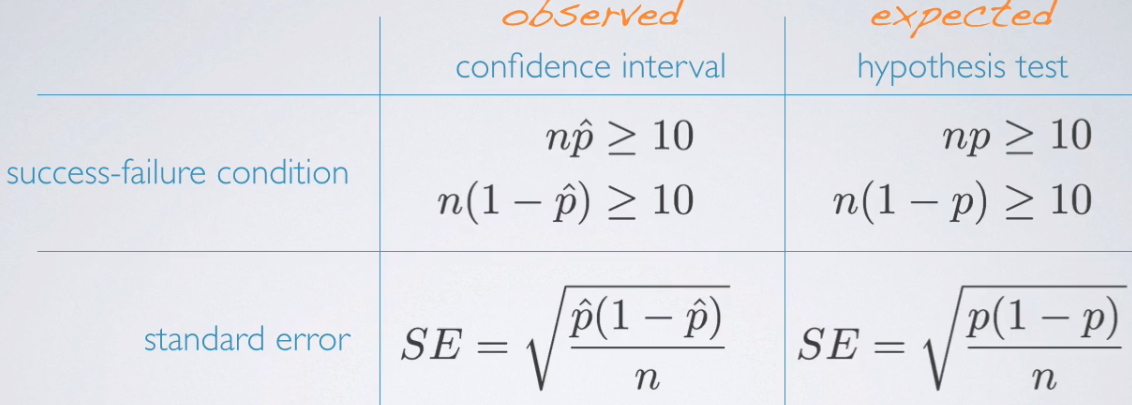
* **Hypothesis testing** comparing 2 = estimate the difference between 2 proportions.
* SurveyUSA poll asked respondents whether any of their children have ever been the victim of bullying, including gender of respondent (parent).
* Distribution of responses by gender of the respondent.



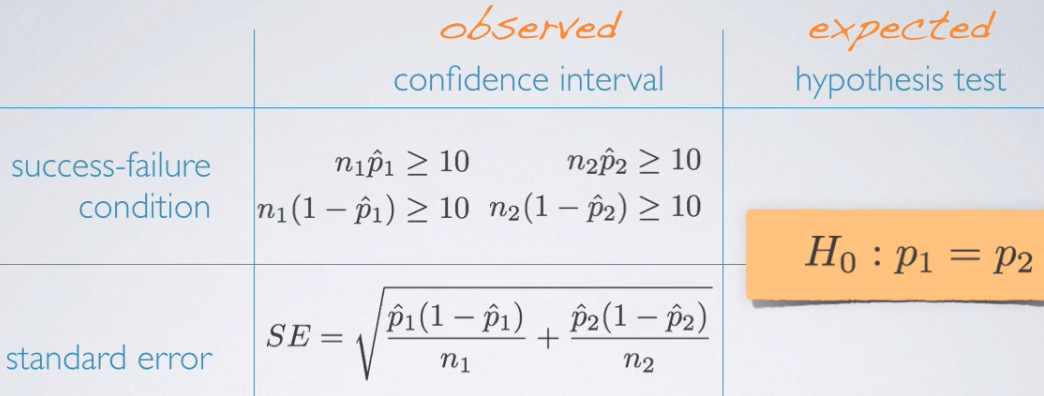
* Remember: We're asking *individuals*, NOT families or households.
* If we see differences between % of bullied kids of males + of females, these may be due to a variety of reasons (single parents, same-sex marriage)
* We're taking the narrow=minded view = 1 mother + 1 father in the household (probably true for majority of the population)
* Could be that 1 gender is more likely than other to even know their kid has been bullied + could also be that 1 gender is more likely than other to actually report this on a survey.

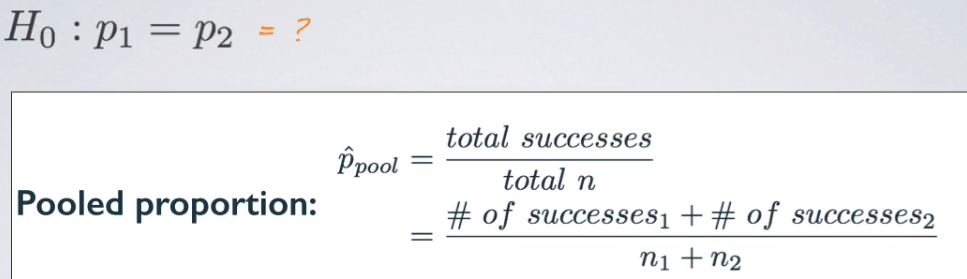
 

* Calculate a test statistic so we can then calculate a p value.
* Working w/ 1 proportion, to check the success/failure condition w/in the context of a CI, we used our *observed* proportions (*observed* successes + failures).
* When doing a hypothesis test, use the value of the *population proportion* we set equal to in the null
* The null value of p is used to calculate the *expected* successes + failures.
* We use p^ in calculation of the standard error for CI vs. using p (null value of population proportion) for the hypothesis test.
* Moral of the story:
* When dealing w/ a **CI**, use **observed** counts + proportions.
* When dealing w/ a **hypothesis test**, use **expected** counts + proportions.



* To translate to working w/ true proportions:
* For CI’s 🡪 look at the total # of observed successes + failures for each group (sample sizes) + multiply them by observed sample proportions to calculate observed # of successes + failures.
* For the calculation of the SE, use observed proportions from the 2 groups as well.
* For HT 🡪 calculating expected successes + failures or expected difference between 2 proportions is not as simple.
* In the null, we say the 2 population proportions should be equal to each other/their difference should be = 0.
* At no point do we define what these p’s should be equal to = don't have a readily-available null value of the population proportion we can use for the 2 groups to calculate expected successes + failures.
* Make one up w/ **pooled proportion**
* Even though the null does not equate the 2 population proportions to something, we could come up w/ a *best guess* for what these could be equal to under the assumption of the null.
* This **pooled proportion** = # of successes / by overall sample size for the 2 groups
* = pooling data from the 2 groups together so it can be calculated as total # of successes in group 1 + total # of successes in group 2 divided by sum of sample sizes for the 2 groups.





> n.male <- 90

> n.female <- 122

>

> suc.male <- 34

> suc.female <- 61

>

> fail.male <- n.male - suc.male

> fail.female <- n.female - suc.female

>

> (p.hat.male <- suc.male / n.male)

[1] 0.3777778

> (p.hat.female <- suc.female / n.female)

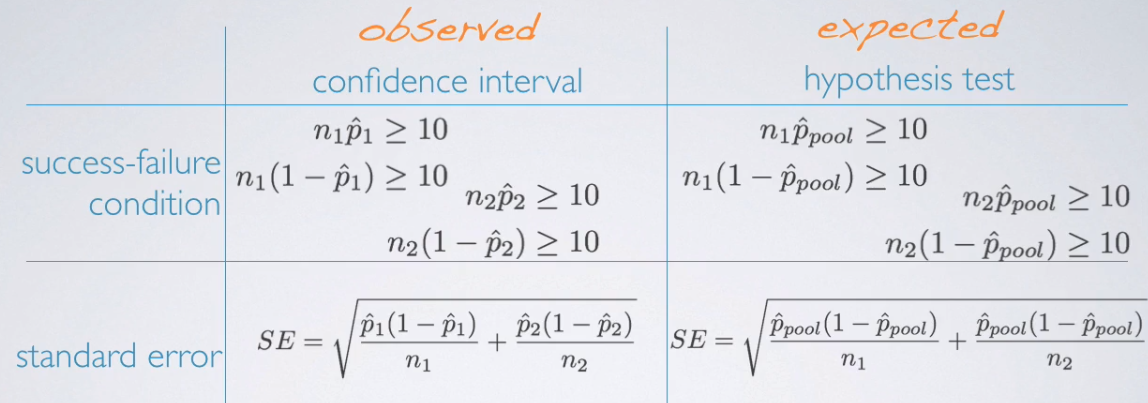
[1] 0.5

>

> (pooled.prop <- (suc.female + suc.male) / (n.male + n.female))

[1] 0.4481132

* Roughly 45% = pooled proportion of males + females who said at least 1 of their children has been a victim of bullying = **good estimate for a common proportion for the 2 groups**,
* For success/failure condition for both groups, use p.hat.pooled to calculate expected # of successes + failures b/c in a hypothesis test, we assume null is true, which states the 2 proportions are equal
* So, we use pooled proportion value as the value they're equal to as the truth in the hypothesis test.
* For SE, plug in the pooled p^ everywhere there was a p^1 or a p^2



* When doing inference for means, we did NOT have different formulas for SE when doing a CI vs. HT
* W/ means our parameter of interest = µ, which is not in the SE formula (S / square root of n)
* doesn't matter what µ is set equal to in the null, it's not going to influence the SE calculation
* **When doing a HT for proportion, we set p to some null value + *that same p* DOES appear in SE calculation**
* **B/c it *does* appear in the SE calculation + b/c we *must assume the null is true* in our calculations, we need to make a different distinction between when we DO have a null we must assume is true (HT) vs. when we DON'T have a null we must assume to be true (CI)**
* Are conditions for inference met for conducting a HT to compare the 2 proportions?
* Condition of independence:
* W/in Group = 90 + 122 = less < 10% of all males + females + we have a random sample, so, sampled males + sampled females can be assumed to be independent of each other
* Between Group = Think about how these data were collected in the 1st place.
* An *overall random* sample + some people happen to be male + some happen to be female.
* Therefore, we really have no reason to expect that sampled males + sampled females in this sample are dependent on each other.
* These are not necessarily **paired** people + even if we had worries about that, the different sample sizes from the 2 groups = they're definitely not 1:1 pairs.
* = No reason to expect dependence between the 2 groups + can assume between group independence condition is met as well.
* Sample size + skew: Remember to consider the success failure condition for a HT for the difference between the 2 proportions = use **pooled proportion** summary table.

> exp.succ.male <- pooled.prop\*n.male

> exp.fail.male <- (1-pooled.prop)\*n.male

>

> exp.succ.female <- pooled.prop\*n.female

> exp.fail.female <- (1-pooled.prop)\*n.female

>

> (exp.succ.male >= 10 & exp.fail.male >= 10)

[1] TRUE

> (exp.succ.female >= 10 & exp.fail.female >= 10)

[1] TRUE

* Can now assume **this sampling distribution of the difference** between the 2 sample proportions is nearly normal.
* Conduct a HT a 5% significance level, evaluating if males + females are equally likely to answer “yes”.
* H0: The 2 proportions are = to each other H1: the 2 are different from each other.
* Ultimate goal of a HT = calculate a p value, but 1st need a test statistic, + for that, we need to figure out our sampling distribution.
* Sampling distribution of the difference between the 2 sample means = nearly normal w/ mean 0 (0 comes from our null value) + we calculate the SE using the pool proportion.

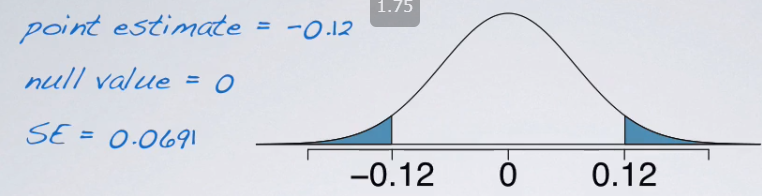
#(z <- (p.hat - h0)/se.prop) # observed - null / n = test statistic

> (se.prop.ht <- sqrt(((pooled.prop\*(1-pooled.prop))/n.male) + ((pooled.prop\*(1-pooled.prop))/n.female)))

[1] 0.06910121

> (point.estimate <- p.hat.male - p.hat.female)

[1] -0.1222222



> h0 <- 0

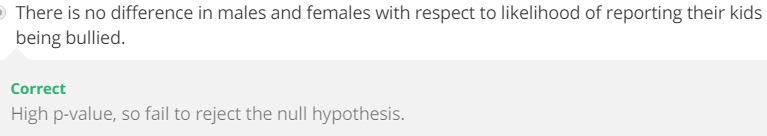
> (z <- (point.estimate - h0)/se.prop.ht) # observed - null / n = test stat

[1] -1.768742

* Our p value = AUC of the absolute value of the z-score beyond 1.74 (-1.74 or lower + 1.74 or higher)

> 2\*pnorm(point.estimate,h0,se.prop.ht)

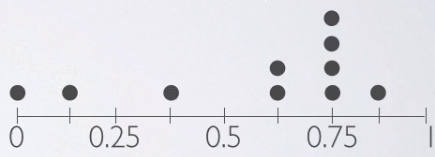
[1] 0.0769369

* Final step = compare this to our significance level, + finally make a decision on the research question we were working w/ 🡪 > sig. level = fail to reject null
* 

***4.4.2 Simulation-Based Inference For Proportions + Chi-Square Testing***

**Small Samples Proportions**

* If success/failure conditions is not met 🡪 **inference via simulation** assuming the null is true.
* B/c if doing any sort of HT where ultimate goal = to get a p-value, the definition of the p-value stays regardless of what type of method you're using.
* p = an observed or more extreme outcome given the null is true.
* Want to make sure that throughout our HT, we act as if that null is true 🡪 set up a **simulation scheme which assumes that null is true**.
* Ex: Paul the Octopus = correctly predicting outcome of soccer games during 2010 World Cup.
* Given 2 boxes, w/ food + flags of the countries playing + predictions = whichever box he chose
* Became famous b/c he predicted all 8 World Cup games correctly.
* *Does this provide convincing evidence Paul actually has psychic powers (does better than randomly guessing)?*
* Because in this setup he had only 2 countries to choose from, if he is randomly guessing, he would be expected to get right 50% of the time
* The null claims he does not have psychic powers + he's simply randomly guessing = set the true proportion of success **H0: p = 0.5.**
* If he's doing better than random guessing, the alternative should say **H1: p > 0.5**
* Sample size = 8 + Paul guessed all correctly 🡪 p^ = 1 or 100%.
* Check conditions for inference
* Independence = can assume his guesses are independent of each other from 1 time to the other
* Sample size/skew = check success/failure condition.
* 8\* null value of .5 = 4 🡺 success/failure is not met = *distribution of sample proportions cannot be assumed to be nearly normal* = we cannot use methods that rely on CLT + the normality of the sampling distribution to find our p-value.
* **Simulation-based inference** comes to the rescue
* Remember ultimate goal of HT 🡺 a p-value = probability of observed or more extreme outcome given the null is true
* Want to devise a simulation scheme that assumes the null is true + repeat the simulation + record the relevant sample statistic many times + finally calculate the p-value as the proportion of simulations that yield a result favorable to the alternative.
* Given our null value = 0.5, we can use a fair coin + label H = successes (correct guesses)
* 1 simulation can be comprised of 8 flips + recording proportion of H
* Trying to simulate whatever Paul did as many times as possible + need to think of his 8 trials as 1 **batch**.
* Want at each simulation to recreate that batch of 8 trials + calculate his rate of success (= 1)
* Try to see *if we leave things up to chance*, what does the rate of success come out to be?
* Repeat simulation many times then, recording proportion of H at each iteration + finally calculate % of simulations where simulated proportion of H is at least as extreme as the usual observation.
* Ex: 1st iteration = sample proportion/proportion of success= 7/8 = 0.875
* Record this #, repeat, + collect #’s on a dot plot
* 10 simulations.



* Observed outcome was 100% success, so the p-value can be defined as: probability of 100% or more success (doesn't make sense, given that the true (expected) rate of success was only 50%)



* We don't have any simulated sample proportions that actually fit the bill, so based on this simulation, our p-value = 0 (almost)
* Chances are if we had actually done this properly w/ about 10k or so simulations, we’d get a small # which would also yield a rejection of the null, but it may not be exactly 0

> source("http://bit.ly/dasi\_inference")

> # create set of the actual observed values

> paul <- factor(c(rep("y",8),rep("n",0)), levels = c("y","n"))

> # perform simulation to estimate a proportion using a hyp. test

> inference(paul, est = "proportion", type = "ht", method = "simulation",

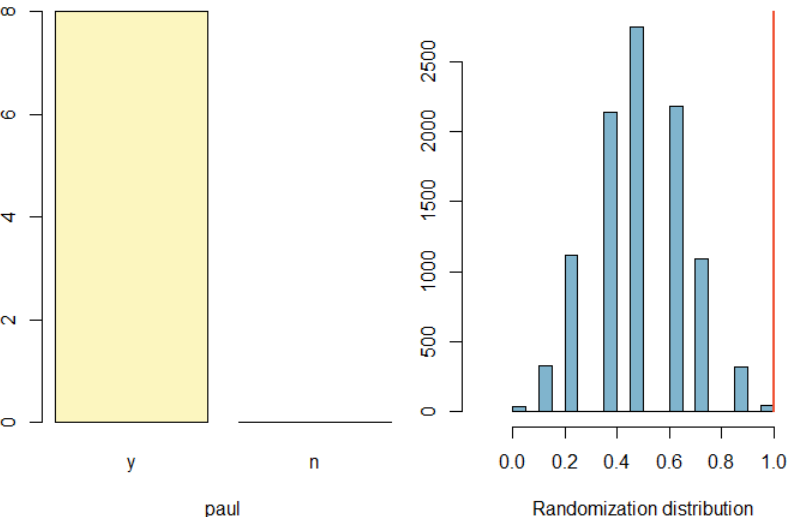
success = "y", null = .5, alternative = "greater")

Summary statistics: p\_hat = 1 ; n = 8

H0: p = 0.5

HA: p > 0.5

p-value = 0.0039



* In this case, the p-value w/ 10k simulations (default for this function) = be 0.0039, meaning again we’d reject the null.
* If Paul was indeed randomly guessing, the probability he’d get all 8 games correct = .0039
* This data provides convincing evidence Paul did better than random guessing
* This does NOT mean we found evidence Paul is psychic + chances are we've made some sort of an error where the null should NOT have been rejected.
* We had a pretty small sample size + it appeared to show a trend in a certain way, + what those particular data yielded was a small p-value based on which we’d definitely reject the null.
* But, we might be making a Type 1 Error = rejecting a null that says Paul simply randomly picks when we shouldn't have.
* The possibility would be to try to collect a little more data from Paul
* English saying = “know something like the back of your hand” = know something very well
* Mythbusters tested the validity of the saying w/ recruited 12 volunteers, each shown 10 pictures of backs of their hands (while wearing gloves so they couldn't actually see their own hands) + were asked to ID their own hand among the 10 pictures.
* 11/12 completed task successfully = were indeed able to recognize the backs of their own hands
* *Do people do better than random guessing when it comes to recognizing the back of their own hand?*
* For each person, they're picking between 10 pictures, so if randomly guessing, probability of success would be = **10% or 0.1 = H0**
* If doing better than random guessing, **probability of success > 0.1 = H1**
* W/ such a small data set, we're don’t meet the success/failure condition = need a simulation-based method to evaluate these hypotheses.
* We want to use a 10-sided fair die to represent the sampling space + want to call a success = guessing correctly + all other outcomes = failures (guessing incorrectly)
* Must assume the null is true + probability of “guessing correctly” from 10-sided fair die = .1, b/c in our null we are setting p = .1
* Then roll the die 12 times, each representing 1 of 12 people in the experiment.
* Then, count # of rolls resulting in 1’s (what we're calling a success) + calculate proportion of correct guesses in 1 simulation of 12 rolls
* Repeat this 100 times, each time recording proportion of simulated success in a series of 12 rolls
* Finally, create a dot plot of the 100 proportions + count # of simulations where proportion is 11/12 (observed outcome)
* Want a p-value for getting 11/12 correct or something more extreme

> source("http://bit.ly/dasi\_inference")

> > # create set of the actual observed values

> hand <- factor(c(rep("y",11),rep("n",1)), levels = c("y","n"))

> > # perform 100 simulations (default) to estimate proportion in a hyp. test

> inference(hand, est = "proportion", type = "ht", method = "simulation",

success = "y", null = .1, alternative = "greater", nsim = 100))

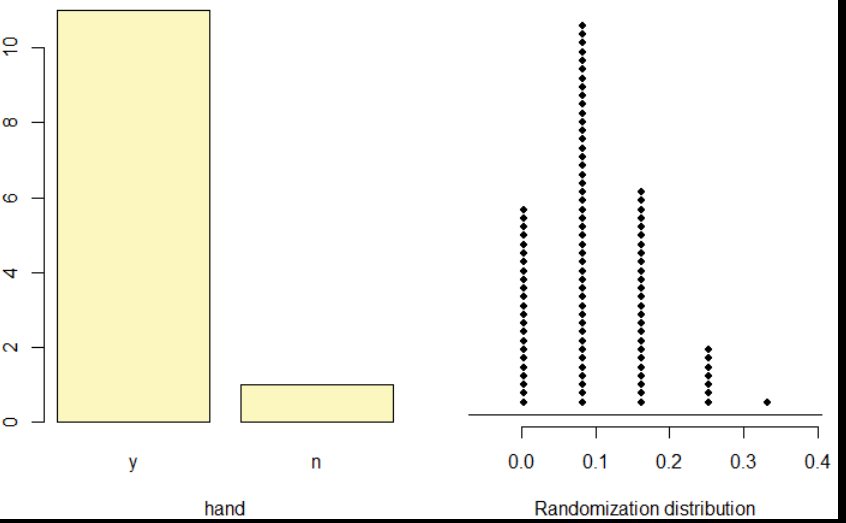
Single proportion -- success: y

Summary statistics: p\_hat = 0.9167 ; n = 12

H0: p = 0.1

HA: p > 0.1

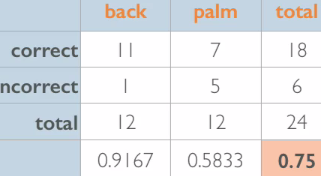
p-value = 0



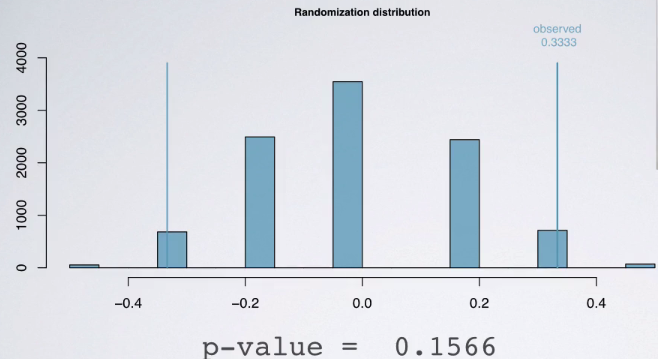
* We get a pretty small P-value + can see from the distribution of the possible simulated proportions that it is quite unlikely to get 11/12 right (p^ = .9167)
* p-value = probability p^ >= .9167 given the true population proportion = .1 🡺 almost 0.
* **There is an almost 0% of 11 or more out of 12 people recognizing the backs of their hands if they were in fact randomly guessing**

**Comparing Two Small Sample Proportions**

* Next we compare 2 proportions coming from 2 groups, both of which are small samples, so our success failure condition is not met.
* Redid the back of hand experiment w/ *palms* of the hands 🡺 only 7/12 subjects were able to ID the palm of their own hand correctly 🡺 success rate = 58.33%
* Do these data provide convincing evidence that there is a difference in how good people are at recognizing the backs + the palms of their hands?
* In this case our null H0: There is no difference between the 2, so p.back – p.palm = 0
* H1: = There is a difference, so p.back – p.palm != 0.
* Check conditions
* W/in group Independence: we can assume guess of 1 subject = independent of another subject
* Just b/c b/c one might be good at guessing doesn't mean another would be
* Between groups Independence: Though we have the same subject + even though we are asking different questions, this is actually not met.
* Assume it to be met for simply illustrative purposes so we can work through how to simulate this scenario.
* Sample size/skew condition: Doing a HT comparing 2 proportions + our null is that 2 proportions are equal to each other.
* To check success/failure condition, we need our **pooled proportion** estimate.



* Overall (*pooled* proportion), of 24 guesses, 18 were correct 🡪 pooled proportion = 0.75.
* Within each group there are 12 guesses, so 75% of 12 = 9 expected # of successes and (1-.85)\*12 = 3 expected # of failures
* Success failure condition is clearly not met + therefore we cannot rely on the sampling distribution to be nearly normal or use any methods that assume so.
* Need to do a simulation test
* W/ 24 trials, we can use 24 index cards for each guess + mark 18 as correct + remaining 6 as wrong.
* what had happened in our actual experiment
* Shuffle the cards + split them into 2 groups of size 12 for back of the hand + palm of hand.
* Then calculate the difference between the proportions of “correct” in the back + palm decks + record this number.
* Repeat this many times to build our random distribution of differences in simulated proportions
* Remember, simulate the experiment *under the assumption of independence*/*leaving things up to chance*.
* If the results from the simulations look like the data, the difference between the 2 proportions of correct guesses can be said to be due to chance.
* If, on the other hand, results from the simulation do NOT look like the data, the difference between the proportions of correct guesses in the 2 groups can be concluded to not be due to chance, but b/c people actually know the backs of their hands better.



* Heights of bars = how many times w/in these 10k simulations a particular simulated p^ was achieved
* Remember, p value = probability of observed or more extreme outcome given the null is true.
* When we think about the *observed*, we want to think about the difference between the success rate in the back of the hand group + the success rate in the palm of the hand group
* This is the corresponding **point estimate** that looks like our null but is based on our sample data.
* The difference between the 2 proportions come out to be ~33%, so the p value = % of simulations >= 33% away from the center of the distribution (always at 0 b/c we assuming the null is true + we're leaving things up to chance)
* W/ a p-value of 0.16, we fail to reject the null + say that there is NOT convincing evidence based on these data that people are better or worse at (or there's some difference between) how they recognize the backs vs. the palms of their hands.

**Chi-Square Goodness of Fit Test**

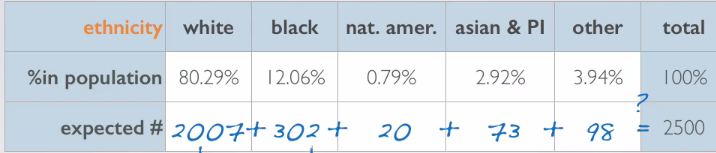
* **Chi-square goodness of fit** = evaluate distribution of 1 categorical w/ > 2 levels = compare distribution of a categorical variable to a hypothetical distribution.
* Ex: In a county where jury selection is supposed to be random, a civil rights group sues the county, claiming racial disparities in jury selection.
* Distribution of ethnicities of people in the county eligible for jury duty based on census results:



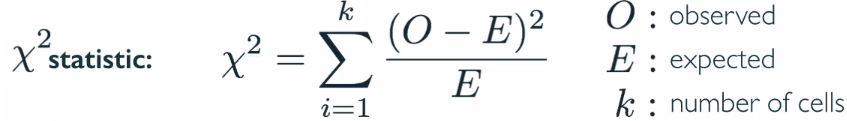
* We are also given the distribution of 2500 people selected for jury duty in the previous year



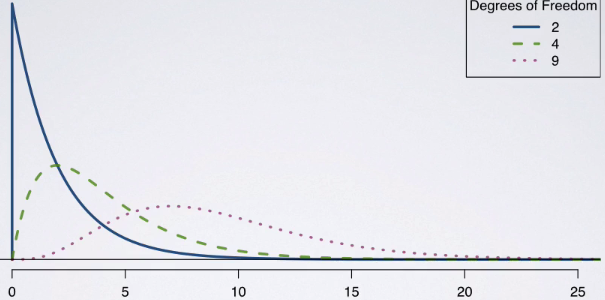
* The court retains you as an independent expert to assess statistical evidence of discrimination.
* Formulate this issue as a hypothesis test w/
* h0 = nothing going on + people selected for jury duty are a simple random sample from the population of potential jurors + the observed counts of jurors from various race ethnicities follow the same ethnicity distribution in the population.
* H1: There is something going on + people selected for jury duty are NOT a simple random sample from the population of potential jurors + observed counts of jurors from various ethnicities do NOT follow the same race ethnicity distribution in the population.
* When we're evaluating the distribution of 1 categorical variable against this hypothetical distribution (true distribution of potential jurors in our population), our null says the observed data follow the hypothesized distribution + the alternative says the observed data do not
* **To evaluate these hypotheses, quantify how different observed counts are from expected counts**
* If observed counts are very different from expected (deviations are large from what is expected based on sampling variation or chance alone), we’d have strong evidence for the alternative.
* This is a **goodness of fit test** = evaluating how well observed data fit an expected distribution.
* If jury selection is random, we’d expect the observed count to follow the % distribution in the population, meaning we’d expect 80.29% of the 2500 people to be white (~ 2007 jurors) if in fact jury selection is random.
* This doesn't mean that's *exactly* what is meant to happen 🡪 If jury selection is random, we’d of course expect some sampling variation or chance around this.
* But we're evaluating if observed counts are so different from expected counts that we suspect something is going on here or if they’re only slightly off from the expected counts



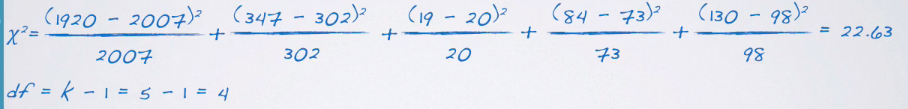
* Once you do this expected calculations, always check to see that the counts calculated actually add up to total sample size.
* May need to do a little rounding to get these counts right
* **Conditions for GOF**
* Independence = make sure that our sampled observations are independent of each other via random sampling or random assignment
* If we're sampling w/out replacement, need sample size to be < 10% of the population
* Also make sure each case only contributes to 1 cell in the table (no juror is black AND white)
* Sample Size = make sure each particular scenario/cell must have at least 5 expected cases.
* Earlier w/ the calculated expected counts in the above table, + it appears we indeed have at least 5 expected cases + also have no reason to believe that these observations are not independent of each other = we have met the conditions for this hypothesis test.
* Need to develop a new test statistic for count data.
* General form of a test statistic = **point estimate – null value / standard error of the point estimate**
* There are 2 things a test statistic tries to accomplish.
* ID’s the difference between a point estimate + expected value assuming the null were true.
* Standardizes that difference using the standard error of the point estimate.
* **Chi-square statistic** = observed counts - expected counts for *each cell*, squared, divided by the expected counts, summed over all cells.



* **Cell** = levels of the categorical variable. + we're introducing a new term here,
* We have 5 cells (ethnicities), so k = five.
* We square the differences between observed + expected counts to make sure our standardized differences are positive, otherwise, positives + negatives will cancel each other out
* Another way of getting rid of negative signs = use an absolute value sign, but by squaring, we accomplish 1 more thing = **highly unusual differences between observed + expected counts appear even more unusual.**
* In order to determine if the calculated chi-squared statistic is “unusually high” or not, we need to 1st describe its distribution, which has only 1 parameter = degrees of freedom.
* dF influences the shape, center, + spread of the chi-square distribution
* For GOF test, the dF = k-1



* More dF = flatter chi-square distribution w/ more spread and greater center
* Null = observed counts of jurors from various race ethnicities follow the same ethnicity distribution in the population.
* Alternative = They do not follow the same race ethnicity distribution in the population.



> ethnicity.perc <- c(.8029,.1206,.0079,.0292,.0394)

> exp.ethnicity.counts <- round(2500\*ethnicity.perc)

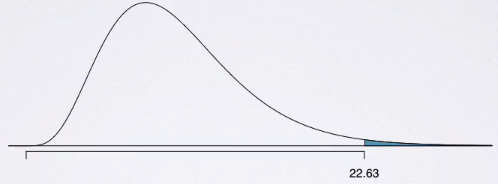
> obs.ethnicity.counts <- c(1920,347,19,84,130)

>

> (chi.sq <- sum((obs.ethnicity.counts - exp.ethnicity.counts)^2/exp.ethnicity.counts))

[1] 22.63311

* In order to find our p-value, we also need to know something about the distribution of the chi-square statistic, + for that we need our dF 🡪 five levels of ethnicity – 1 = 4
* The p-value for a chi-square test = tail area above the calculated test statistic
* Just like F-tests
* B/c the test statistic is always positive, + a higher test statistic = a higher deviation from the null.



* It's right skewed + it's a squared value, so it’s always a positive # + we shade the tail area beyond the observed chi-square statistic we calculated.

> dF <- length(ethnicity.perc)-1

> pchisq(chi.sq,dF,lower.tail = F)

[1] 0.0001498896

* W/ such a small p-value, we’d reject the null = the data provide convincing evidence that the observed distribution of the counts of race ethnicities of jurors does NOT follow the distribution in the population.