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

***COURSE 2 - Inference***

**WEEK 2- Inference and Significance**

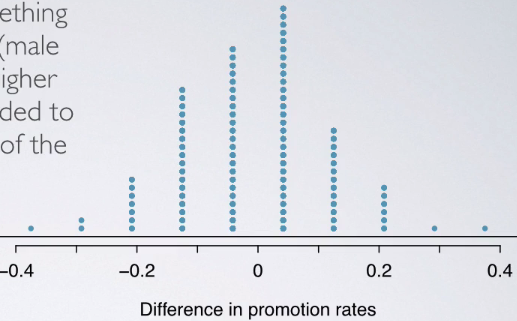
***4.2 Hypothesis Testing***

**Another Introduction to Inference**

* Remember:

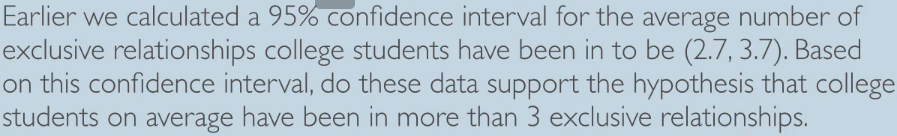


* Experiment done on male bank managers presented the same employee file/resume w/ all qualifications being the same as original but ½ were labeled male + ½ were labeled female.
* In EDA, the % of males promoted were 21/24 = 88% of males.
* However only 14/24 = 58% of females were promoted.
* We were clearly able to see a difference between % of males and females promoted.
* However, instead of jumping the gun, we said that there could be 2 competing claims that actually explain what's going on here:
* 1) Nothing 🡪 promotion + gender = independent 🡺 observed difference in proportions due to chance = **Null**
* 2) Something is going on 🡪 promotion + gender = dependent 🡺 observed difference in proportions due to gender discrimination (not due to chance) = **Alternative**
* Did a simulation-based inference under the assumption of the null being true/assumption of independence (leaving everything up to chance)
* For each simulation, we recorded this observed difference in promotion proportions:
* In our original observed data, this was 88 – 58 = 30%.
* We looked to see if 30% was an unusual outcome for this difference



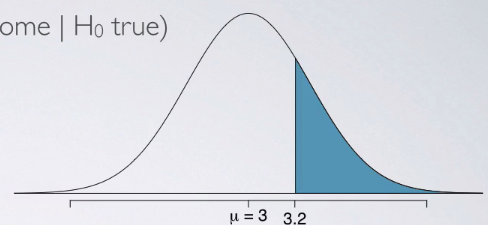
* The actual data value of 30% seems very unlikely based on the dot plot above 🡪 reject the null in factor of the alternative
* Framework 🡺 Start w/ Null H0 (status quo) 🡪 Create alternative H1 to represent research question (what we’re testing for) 🡪 conduct hypothesis test under assumption the null is true either via simulation or theoretical methods that rely on the CLT)
* If test results suggest the data do NOT provide convincing evidence for the alternative, stick w/ the null 🡺 If they do, reject the null in favor of the alternative.

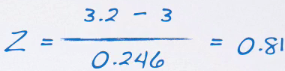
**Hypothesis Testing (for a Mean)**

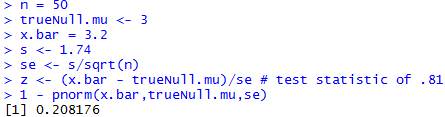
* **Null** = H0 = skeptical perspective or claim to be tested set to a parameter value
* **Alternative** = H1 = claim under consideration often represented by a range of possible parameter values (<, >, or != the null)
* Skeptics will not abandon H0 unless evidence in favor of H1 is *so strong* we can reject H0 in favor of H1
* 
* H0: μ = 3 (students have been in 3 exclusive relationships on average)
* H1: μ > 3 (students have been in > 3 exclusive relationships on average)
* Our null is included in our CI, so therefore μ *could* = 3
* CI says any value w/in it is the possible true population mean, so therefore 3 is a possible solution so we cannot reject H0 in favor of H1
* But this doesn’t tell us the p-value (likelihood of certain outcomes under the null), based on which we can make decision on the hypotheses
* \*\*Note: Hypotheses are always about **population parameters**
* Remember p = the probability of an observed (or more extreme) outcome value given H0 is true
* Here, this is x.bar > 3.2 | H0: μ = 3
* Since we assume H0 is true, we can use it to construct the sampling distribution based on the CLT

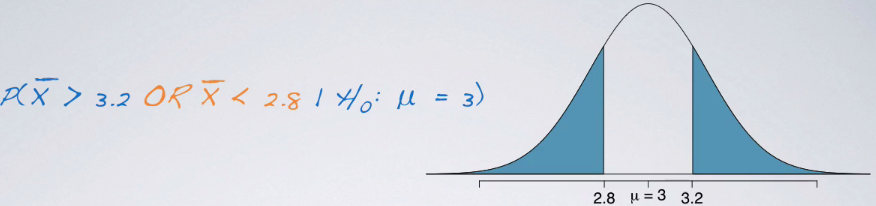


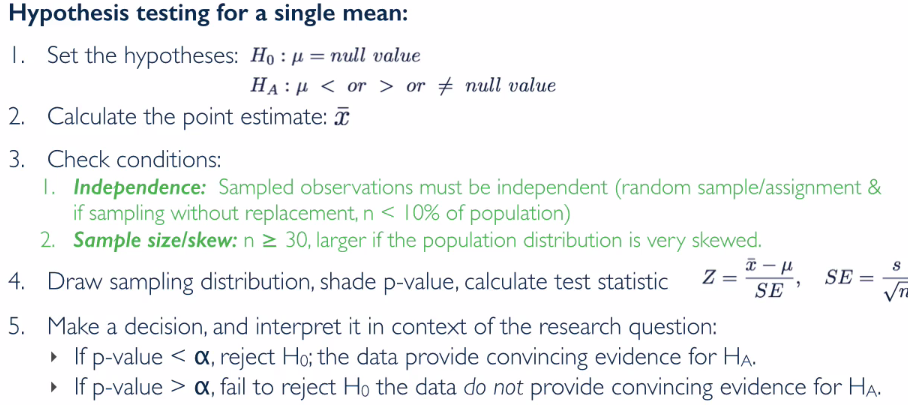


* X.bar is nearly normally distributed w/ mean = 3 + SE = .246
* Now draw the curve + shade area of interest corresponding to the p-value
* 



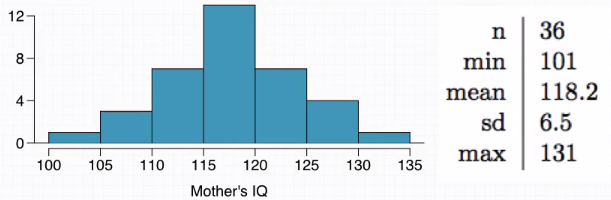


* So, **p = .209**
* We used the **test statistic** z to calculate **p-value** = *probability of observed data at least as favorable to the alternative as our current dataset, if the null were true*
* If p is low (< significance level **α** usually set at 5% or .05), we say it’s very unlikely to get the observed data if the null were true 🡪 reject it
* If p is high (> significance level **α** usually set at 5% or .05), we say it’s likely to get the observed data if the null were true 🡪 fail to reject it
* Our p-value is high (> .05) 🡺 even though we have a sample mean x.bar > 3, there is not enough evidence to reject the null that the population average # of exclusive relationships for college students = 3 🡪
* If in fact college students have been in = 3 exclusive relationships on average, the p-value tells us there is a 21% chance a random sample of 50 students would yield a sample mean >= 3.2
* Since this is a pretty high probability, we’d think a sample mean >= 3.2 is likely to happen simply by chance
* We fail to reject the null + these data do not provide convincing evidence college students have been in > 3 relationships on average + the difference between the null of 3 relationships + the observed sample mean of 3.2 relationships is simply due to chance or **sampling variability**.
* Often, instead of looking for a divergence from the null in a *specific direction* (< or >), we are interested in divergence in any direction 🡪 **two-sided/tailed hypothesis tests**,
* Definition of a p-value = the same regardless of doing a 1 or a 2-sided test.
* Calculation becomes slightly different + ever so slightly more complicated 🡪 need to consider at least as extreme as the observed outcome in BOTH directions away from the mean.
* Now set p-value now to be x.bar > 3.2 OR x.bar < 2.8, given the null is true = population mean = 3
* *Use 2.8 b/c it’s the same distance from 3 as 3.2*
* 
* Already know the upper tail = 0.209 + since this is a symmetric distribution, lower tail also = 0.209
* Therefore p-value = probability on upper tail + probability on the lower tail = ~41.8%.



**Hypothesis Testing (for a Mean) Examples**

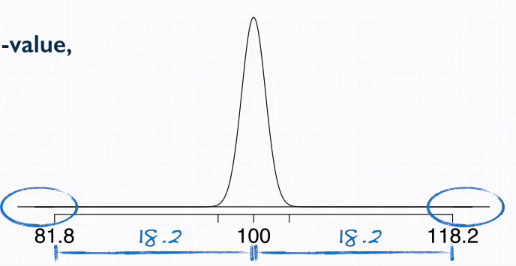
* Researchers investigating characteristics of gifted children collected data from schools in a large city on a random sample of 36 children IDed as gifted children soon after they reached the age of 4.
* In this study, along w/ variables on the children, researchers also collected data on mothers' IQ scores
* The histogram shows the distribution of these data, and also provided our some sample statistics.



* We're asked to perform a hypothesis test to evaluate if these data provide convincing evidence of a difference between average IQ score of mothers of gifted children vs. average IQ score for the population at large, which happens to be 100. Use a significance level of .01
* 1st, define parameter of interest (mean of IQ scores of mothers of gifted children) and define the hypothesis: 
* 2nd, define point estimate 🡪 given to us as the sample mean 118.2 (average IQ scores of the 36 mothers)
* 3rd, check conditions
* we have a random sample + 36 children total is < 10% of all gifted children, therefore **we have independence**
* can assume IQ score of 1 mother in the sample is independent of another’s
* n > 30 and the histogram shows that distribution is not skewed, so we have **nearly normal sampling distribution**



* 4th, draw curve, shade p-value, calculate test statistic (z)



* Lower tail is equidistant from the population mean as the given sample statistic
* See p-value regions are very small

> n <- 36

> min <- 101

> x.bar <- 118.2

> trueNull.mu <- 100

> sigma <- 6.5

> max <- 135

> N <- 100

> alpha <- .01

>

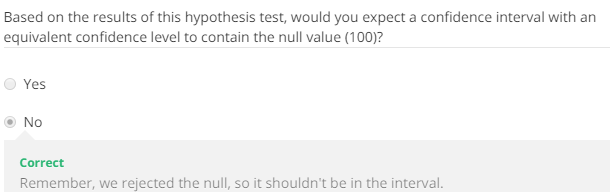
> se <- sigma/sqrt(n)

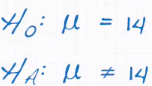
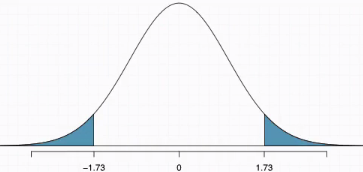
> (z <- (x.bar - trueNull.mu)/se) # test statistic of .81

[1] 16.8

> 1 - pnorm(x.bar,trueNull.mu,se)

[1] 0

* Such a high z-score = our observation is usually far from the center = **the p-value is ~0 = strong evidence against the null = reject null**
* We conclude the data provide convincing evidence of a difference between the average IQ score of mothers of gifted children and the average IQ score for the population at large.
* 
* A stats student interested in sleep habits of domestic cats took a random sample of 144 cats + monitored their sleep.
* The cats slept an average of 16 hours/day +, according to online resources, domestic dogs sleep on average 14 hours/day.
* We want to find out if these data provide convincing evidence of different sleeping habits for domestic cats + dogs w/ respect to how much they sleep.
* The test statistic calculated was 1.73.

> z <- 1.73

> n <- 16

> x.bar <- 16

> trueNull.mu <- 14

> alpha <- .01

> (se <- (x.bar - trueNull.mu)/z)

[1] 1.156069

> 1 - pnorm(x.bar,trueNull.mu,se)

[1] 0.04181514

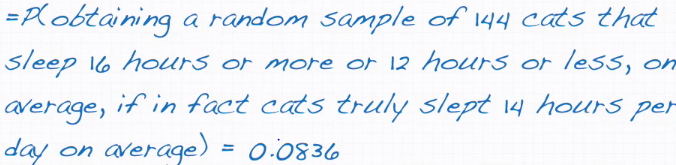
* This is the p-value for ONE of the tails, so our total p-value is

> 2\*(1 - pnorm(x.bar,trueNull.mu,se)) # two-tailed test

[1] 0.08363028

* What is the interpretation of this p-value in context of these data?
* **p-value** = probability of observed/more extreme outcome *given the null is true*.
* An “*observed*” outcome of 16 hours per day
* Something more extreme could be in the positive (> 16) or negative direction (< 16), b/c our alternative is 2-sided.
* To figure out the cutoff value for the *negative* side of the distribution is going to be, we need to venture an equal distance from the mean as our sample statistic 🡪 16 – 14 = 2 hour difference.
* So, we’d do 14 – 2 = 12 hours or lower is also of interest to us.



* Saying the null is true 🡪 means mu = 14, or, in other words, that cats do truthfully sleep 14 hours on average.
* So, putting all of this together:
* 
* According to some guidelines we've been using so far, that would be considered kind of high

***Significance***

**Inference for Other Estimators**

* So far, been focusing attention on doing inference for ONE population mean w/ data from ONE sample to find the sample mean + sample SD + then use those values to say something meaningful about the unknown population mean.
* Can actually use the same methods for doing inference for a variety of other estimators = unified nature of hypothesis testing + CI’s.
* Methods for doing hypothesis tests and constructing CI’s, can be easily adapted for any estimator that has a **nearly normal sampling distribution** 🡪 a sample mean **x.bar**, or a difference between 2 sample means x**.bar1 – x.bar2** (useful for comparing 2 groups of population means)
* Another estimator = **sample proportion** **p^**
* Sampling distribution of the sample proportion will also be nearly normal if sample size is high.
* Can then apply the same techniques of inference for proportion or looking at difference between 2 proportions **p^1 - p^2**
* This again gives an avenue by which to compare 2 groups/populations to each other
* Important Assumption about point estimates = *they're unbiased.*
* In other words, the sampling distribution of the estimate = centered at the true population parameter it estimates.
* That is, an unbiased estimate does NOT naturally over/underestimate the parameter, but instead provides a good estimate.
* Sample mean = example of an unbiased point estimate 🡪 CLT says the sampling distribution of sample means is going to be nearly normal, centered at the true population mean.
* W/ a unbiased point estimate w/ a nearly normal sampling distribution, can construct a CI around it.
* *Always start with a point estimate when working with Cis* +then we add + subtract the same amount (**margin of error = critical value \* standard error)**
* Once you have this general structure set up for the CI, just swap out the formula for SE for a different estimate + keep everything else the same.
* A 2010 Pew Research foundation poll indicates among 1,099 college graduates, 33% watch the Daily Show, an American late-night TV Show. The standard error of this estimate = 0.014. Estimate the 95% CI for the proportion of college graduates who watch The Daily Show.
* Given n = 1099, x.bar = .33, se = .014

> n <- 1099

> x.bar <- .33

> alpha <- .05

> se.x <- .014

> (z.crit <- qnorm(1-(1 - .95)/2)) # ~ 1.96

[1] 1.959964

> mOe <- z.crit\*se.x

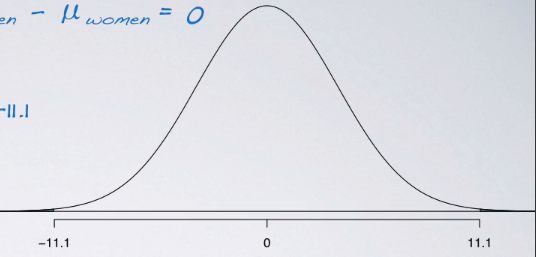
> (ci.low <- x.bar - mOe)

[1] 0.3025605

> (ci.low <- x.bar + mOe)

[1] 0.3574395

* The generic formula for a **CI** for any estimator **= a point estimate +/- a margin of error**
* In this case, our point estimate is a p^, + we have +/- a critical value, **z\***, times our SE (margin of error)
* ***We are 95% confident between 30.3% and 35.7% of college graduates watch the Daily Show***
* Just like with CIs, we can apply the same framework for hypothesis testing to other estimators as well, *as long as the estimator is unbiased and has a nearly normal sampling distribution*.
* If that's the case, we can use the z statistic as our test statistic which is always calculated as a point estimate - the null value (or observed - the mean), divided by some SE.
* The 3rd National Health and Nutrition Examination Survey (NHANES), collected body fat % + gender data from over 13,000 subjects ages 20-80. Average BF% for the 6,580 men in the sample = 23.9 + this was 35% for the 7,021 women. SE for the difference *between* the average male + female BF% = 0.114. Do these data provide convincing evidence that men and women have different average BF%’s? Assume the distribution of the point estimate is nearly normal.



> x.bar.m <- .239

> x.bar.f <- .35

> (x.bar.diff <- x.bar.m - x.bar.f)

[1] -0.111

> se <- .114

> # null says difference = 0

> (z <- (x.bar.diff - 0)/se)

[1] -0.9736842

> (1 - pnorm(x.bar,0,se))

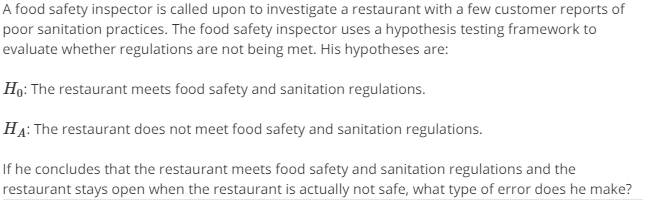
[1] 0.001897383

* W/ such a huge z score, we could tell the p value is bound to be really tiny = reject the null
* In context, we’ve determined that these data provide convincing evidence that the average body fat percentages of men and women are indeed different from each other.

**Decision Errors**

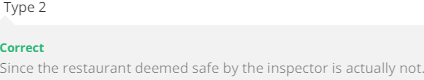
* Hypothesis tests are NOT flawless
* We almost don't know whether the null or the alternative is true 🡪 make decisions based on the **evidence** collected or on the **statistical significance of that evidence**.
* If the null hypothesis is indeed true + you fail to reject it = CORRECT
* If the alternative is true + you reject the null in favor of the alternative = CORRECT
* **Type I error** = rejecting the null when it’s actually true.
* **Type II error** = failing to reject the null hypothesis when the alternative is true
* We need to consider all possibilities.



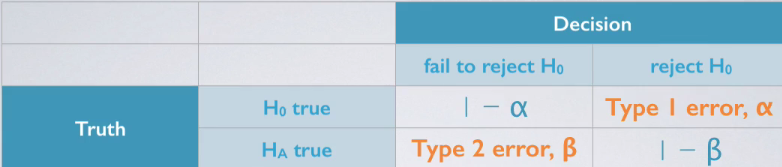
* The null says that the “defendant” is innocent (until proven guilty) = H0
* The alternative says the defendant is guilty = H1
* Declaring the defendant innocent when they are actually guilty = Type II (failed to reject the null in favor of the alternative when the alternative is actually true
* Declaring the defendant guilty when they're actually innocent = Type I (rejected the null in favor of the alternative when the null was true)
* *Which error is the worse to make?* ***🡪*** **this a subjective question.**
* One can make the argument, however, that there's a consensus around a quote from William Blackstone = “Better that 10 guilty persons escape than 1 innocent suffer”.
* If you agree w/ this when setting up a hypothesis testing scheme, you want to **minimize Type I (False Positive) error rate** of declaring a defendant guilty when they're actually innocent.
* If you do NOT agree w/ this, you think that a **type II (False Negative) error** is a worse offense
* As a general rule, we reject the null hypothesis when p < .05 (set significance level/alpha = .05)
* This means that, **for cases where the null is actually true, we do not want to incorrectly reject it more than 5% of those times**.
* In other words, **when using a 5% significance level, there is about a 5% chance of making a type one error if the null is true.**
* Can simplify this to read as that **alpha = probability of a type I error, given the null is true**
* This is why we prefer small values of alpha b/c if we increase it, it might *sound* good b/c we’d reject our null more often (usually what we’re trying to do)
* However, we’d also *be inflating our type I error rate*.
* **Increasing alpha increases the type I (FP) error rate**
* Choosing alpha is a balancing act. If type I error is dangerous/especially costly, we might want to choose a small significance level, something even smaller than 5%
* Goal in this case = to be very cautious about rejecting the null 🡪 so we demand very strong evidence favoring the alternative before we’d do so.
* If, on the other hand, the type II (FN) error is relatively more dangerous/costly, we want to choose a *higher* significance level
* Increasing significance levels decreases our type II (FN) error rate.
* Goal in this case = be cautious about *failing* to reject the null when the null is actually false.
* 



* 



* **Type I (FP) error** = *rejecting* the null when you shouldn't have w/ the probability of doing = alpha
* **Type II (FN) error =** *failing* to reject the null when you should have w/ probability of doing so is = **beta** (not a trivial calculation)
* The **power** of a test has positive connotation = **probability of *correctly* rejecting the null = 1 – beta**

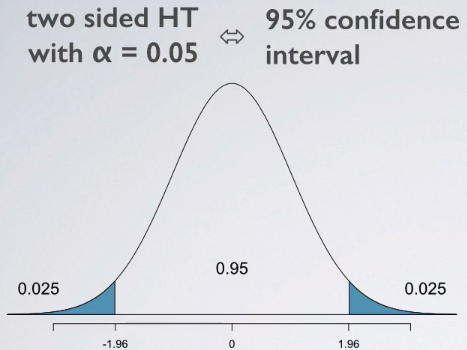


* If the alternative is true, there are 1 of 2 things you can do
* Commit a type II (FN) error w/ probability of doing so = beta
* Correctly reject the null in favor of the alternative w/ probability of doing so = complement of beta 🡪 (1 – beta) = **power**.
* General Hypothesis Testing Goal = **keep both alpha and beta low at the same time**.
* But we know as we push 1 down, the other shoots up 🡪 Want to find a delicate balance between these 2 probabilities.
* If the alternative is actually true, what is the chance that we make a type II error = what is the chance we *fail* to reject the null, even when we *should* reject it? = **beta**
* The answer to this is not obvious.
* If the true population average is *very close* to the null value, it will be very difficult to detect a difference + to reject the null.
* In other words, if the true population average is *very different* from the null value, it will be much easier to detect a difference.
* Clearly then, beta (probability of making a type II error) depends on our **effect size** = the difference between + point estimate + the null value.

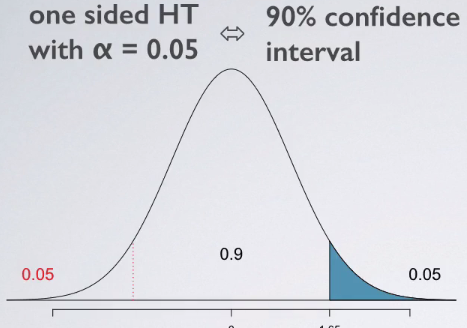
**Significance vs. Confidence Level**

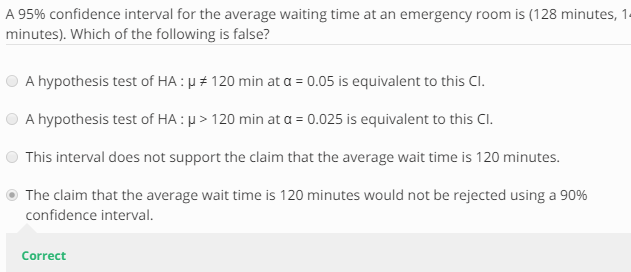
* Results of CI’s and hypothesis tests agree w/ other if we use equivalent levels of **significance + confidence,** which are complements 🡪 5% significance level = 95% CI

BUT whether this compliment rule works or not depends on if doing a 1 or 2-sided hypothesis test



* 2-tailed 🡪 alpha of 0.05 = two 0.025 tail areas
* CI = we're always thinking about the middle % of a distribution
* 2-sided hypothesis test w/ alpha = 0.05 + the 2 tail areas add up to 0.05 🡪 indeed equivalent to a 95% CI.
* If we add up the 95% in the middle w/ the alpha on the 2 tails = 1



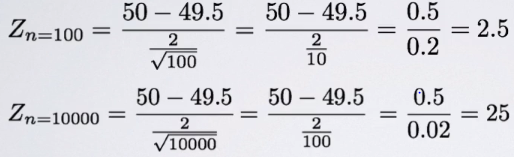
* Looking for where p-value could be anything up to 5% only in 1 tail = could still reject the null
* Since confidences always have to be symmetric + the **confidence level** is always the middle % of a distribution + we can’t have a CI that only goes a certain amount of distance in *one* direction but *more* in the *other* direction, we need to think about lower-tail 5% even though we are not going to use it for any sort of decision making w/in the hypothesis test.
* So w/ a 1-sided hypothesis test w/ alpha = 0.05, your equivalent confidence level = 90% b/c we're allowing for the 5% at the one end you're interested in + must take note of 5% at the other end that you're not interested in so that the CI can be symmetric.
* So w/ a 1-sided hypothesis test, we can't say the significance level + confidence level are complements of each other *directly*.
* We can still use that idea though to figure out how to get from one to the other = the key = always draw your curve + mark what you're interested in
* Oftentimes, you may want to use both methods when doing inference + it doesn't make sense for the methods to not agree with each other.
* So, for the most part, if a given significance level + have accurately determined the equivalent confidence level, the results from the 2 approaches should always equal agree w/ each other
* **2-sided hypothesis w/ a threshold of alpha = a confidence interval of 1 - alpha.**
* **1-sided hypothesis w/ a threshold of alpha = a confidence interval with 1 - 2\*alpha.**
* **If the null is rejected, a CI that agrees w/ the result of the hypothesis test should NOT include the null value.**
* This is b/ ci you're saying you're *rejecting* the null value, so that value should NOT be in the confidence interval
* Otherwise, you’d be contradicting yourself by saying it's a plausible value for the parameter of interest.
* **If we FAIL to reject the null, a CI that agrees w/ the result of the hypothesis test SHOULD include the null value.**
* 

**Statistical vs. Practical Significance**

* Ex: All else held equal, will a p-value be lower if n = 100 or n = 10,000 🡪 2 separate hypothesis tests w/ the same null + alternative hypotheses, sample means = exactly the same, null values = the same, b/c they're driven by the hypotheses + SSs = exactly the same as well, but sample sizes are different

If the sample size is high, **standard error** will decrease b/c square root of n increases

* SE w/in a hypothesis testing framework shows up in the denominator of a test statistic.
* If your denominator goes down, your test statistic is going to go up + closer to the ends of the tails of a distribution as opposed to closer to the center
* If cutoff values/test statistics are actually closer to the ends of tails, then the **p-values** (those tail areas) are going to be getting smaller + smaller.
* **Increase n 🡪 SE error decreases 🡪 test statistic increases 🡪 p-values decreasing**



* We know (using the 68, 95, 99.7% rule) almost all observations under the normal curve lie w/in 3 SD’s the mean, so a z-score of 25 basically means almost no p-value = **highly *statistically* significant**
* However, is it ***practically* significant**?
* **Practical significance =** focus on the **effect size** = difference between point estimate + null value.
* This would be the numerator in our test statistic calculation
* In both instances above, we have the same exact effect size + it's a small effect size

Even w/ a small effect size, which may not be practically significant, we’re able to find a statistically significant result simply by inflating sample size (*something the researcher has control over*)

* **So, when you see highly statistically significant results, make sure you have a critical eye + that you inquire whether effect size is reported + what the sample size is**
* If reporting highly statistically significant results, it's a good idea to let readers know effect + sample size so the discussion is clear if a statistically significant finding is *also* practically significant or not
* **Real differences between the point estimate + the null value are easier to detect w/ large samples**
* **However, very large samples result in statistical significance even for tiny differences between the sample mean + the null value (effect size) even when the difference is not practically significant.**
* In order to make sure findings don't suffer from this problem of being statistically but not practically significant, oftentimes we do some **a priori analysis** before actually doing data collection to figure out (based on characteristics of the variable you’re studying, how many observations to collect)
* Highly recommended that researchers either should do this themselves or consult w/ statisticians, to figure out how many observations to sample before they actually go into doing that.
* Last thing you want to do is collect some data + either don't have enough or have too many observations.
* “To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination. He may be able to say what the experiment died of” - R.A. Fisher