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

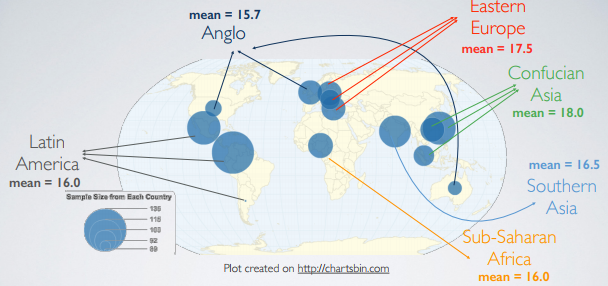
***COURSE 2 - Inference***

**WEEK 3 - Inference for Comparing Means**

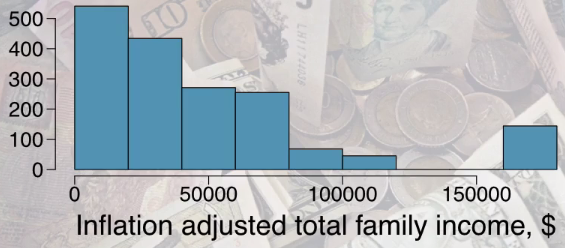
***4.3.1 t-distribution and Comparing 2 Means***

**Introduction**

* Acceptability of Workplace Bullying = study that explores relationship between culture + acceptability of workplace bullying across the globe.
* Researchers collected data using a survey from 1484 alumni + current MBA students from 14 counties on 6 continents + asked some questions on acceptability of **work related bullying**
* **Work related bullying** = giving tasks w/ unreasonable deadlines or exposing workers to an unreasonable workload, so on + so forth.



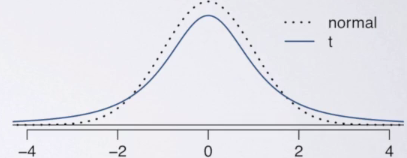
* See a geographic distribution of countries included in the study w/ sizes of circles = how large sample sizes (SS) are from each country.
* SS’s are somewhat consistent across globe + it seems like a pretty even geographic distribution
* Study further categorizes 14 countries into 6 continents + those are the 6 groups we're considering.
* We calculate mean acceptability of work related bullying score for each group (low score = bullying is unacceptable in the workplace, high score = is actually acceptable)
* Can see that the average acceptability is higher in Asia + lowest in Anglo countries.
* But just looking at sample statistics = not possible to determine if differences we're observing are **statistically significant**.
* Want to compare many means to each other
* Look at distribution of inflation-adjusted total family income in the US from a random sample of Americans collected as part of the General Social Survey in 2012



* Distribution is, as expected, pretty right-skewed.
* Suppose we‘d like to estimate typical total family income in the US.
* The CLT provided the basis for constructing a confidence interval for the mean, but what if we're not interested in the mean, ***but the median***?
* *No CLT for the median*.
* New technique for creating CI’s = **bootstrapping =** accomplishing an impossible task = a simulation-based method that doesn't have *as* rigid conditions as the CLT + therefore also works for many estimates beyond the mean

**t-distribution**

* **t-distribution** = useful for describing the distribution of a sample mean when *population SD, sigma, is unknown (almost always)*
* Remember, what purpose does a large sample serve?
* As long as observations are independent + the populations distribution is *not* extremely skewed, a large sample ensures you have a nearly normal sampling distribution of the mean + that the estimate of the **standard error** (SE = S / sqrt(n), *best estimate for unknown pop. SD*) is reliable
* So, if the sample size n is large enough, chances are SE (s) is indeed a good estimate for sigma, + therefore your overall SE estimate is reliable.
* In the age of “big data “why are we talking about small samples.
* It’s true in certain disciplines (especially w/ automatically-recorded data like webpage clicks or Twitter streams), small sample sizes might be irrelevant.
* However, there are disciplines where this is not the case (lab experiment or a study that follows a near-extinct mammal species).
* WE need methods that work well for BOTH large + small samples.
* Uncertainty of the SE estimate = addressed by using the **t-distribution** = also has a bell shape (unimodal + symmetric) + looks a lot like the normal distribution but w/ thicker tails



* Peak of t-distribution doesn't go as high as normal distribution = *t-distribution is somewhat squished in the middle + additional area is added to the tails.*
* This means, under the t-distribution
* observations = *more likely to fall 2 SDs away from the mean than under the normal distribution*
* CI’s constructed using a t-distribution = wider/more conservative than those constructed w/ the normal distribution
* Thick tails = helpful for mitigating the effect of a less-reliable estimate for the SE of the sampling distribution caused by using the sample SD instead of the population SD in its calculation.
* t-distribution (like the standard normal) = always centered at 0 + has 1 parameter = **Degrees of freedom** = determines thickness of the tails.
* In contrast, the normal distribution has 2 parameters 🡪 mean + SD.
* As dF increases, the shape of the t distribution increases + approaches the normal distribution



* We **use the t distribution for inference on a single mean** or **for comparing 2 means when population SDs are unknown (basically always)**
* Calculate t statistic T just like a Z statistic + find the p-value = probability of observed or more extreme outcome values given the null is true (same definition as before)



* Calculate
* probability the absolute value of Z is greater than 2, which is .0455 B
* probability the absolute value of t w/ 50 dF freedom > 2
* Remember t = thicker tails + higher % of observations falling further than 2 SDs from mean
* We're starting to see the effect
* probability the absolute value of t w/ 10 dF freedom > 2

> (pnorm(2,0,lower.tail = F)\*2) # only 1-sided hypothesis

[1] 0.04550026

> pt(2, 50, lower.tail = F)\*2

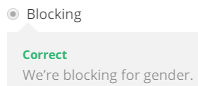
[1] 0.05094707

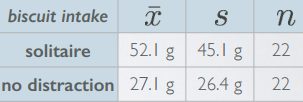
> pt(2, 10, lower.tail = F)\*2

[1] 0.07338803

* So, **as we go from the normal to a t distribution w/ a somewhat high dF to a t distribution w/ low dF, the probability of the test statistic being more than 2 SDs away from the mean increases.**
* Suppose you have a 2-sided hypothesis test + your test statistic = 2.
* Under which of the above scenarios would you be able to reject the null at the 5% significance level?
* 1st scenario = p = 4.55% which is < 5% = reject the null (barely)
* 2nd = p > .05 = fail to reject the null (barely)
* Last scenario = definitely fail to reject the null.
* As we get more conservative w/ a t distribution (lower dF = wider CI’s), we also become less likely to be able to reject the null (more likely to have it in the CI)
* Generally, dF is tied to sample size 🡪 if n is low, it is not as easy to reject the null + stronger evidence is needed in order to be able to do so.
* *This* t-distribution = **student's t distribution 🡪** William Gosset = head experimental brewer at Guinness in early 1900's w/ main role = to experimentally brew + gradually improve a consistent + economical barrel of the Guinness stout.
* This required sometimes working w/ small samples b/c maybe he’d just have few batches to try
* So, much development of the t-distribution comes from trying to make Guinness taste better
* Since Guinness was worried about trade secrets getting out, Gosset was asked to publish any work he was doing under a pseudonym and “Student” was the name that he chose for.
* While others, like Fisher, continued to work on the t-distribution, even Gosset's own foundational work, the distribution is still named after his pseudonym

**Inference for a Single Mean**

* Study = Playing A CPU Game During Lunch Effects Fullness, Memory For Lunch + Later Snack Intake.
* In this study, researchers evaluated the relationship between being distracted + recall of food consumed + snacking, w/ the idea that if you're distracted while you're eating, you may not remember what you eat.
* They also hypothesized failure to recall food consumed might lead to increased snacking later on.
* Sample = study consisted of 44 volunteer patients, half men, half women, who were randomized into 2 groups, 1 asked to play solitaire on the CPU while eating + to win as many games as possible, + the other group was asked to eat lunch w/out any distractions, focusing on what they're eating + thinking about the taste of the food + that they're eating.
* 
* 
* Both groups were provided the same amount of lunch + afterwards, while they were waiting around, they were offered biscuits to snack on.
* Researchers measured how many biscuits subjects consumed



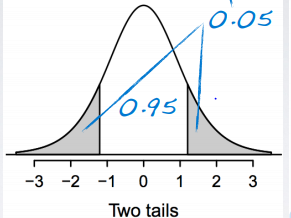
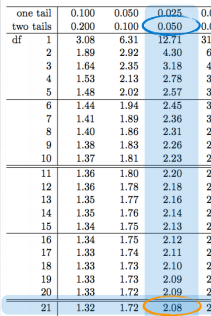
* This summary statistics suggest distracted eating groups snack more after lunch (x.bar.s = 52.1 g of biscuits compared to 27.1 g (x.bar.n) for the other group.
* We're also given the SDs for both groups, as well as the sample sizes, n, both = 22
* Goal = estimate average snacking level for distracted eaters.
* Estimating a population parameter entails a **CI = point estimate +/- a margin of error**.
* **Margin of error** = a critical value \* the standard error
* Since we're doing **inference on the mean** = use **t statistic**
* This SE of x.bar = s / sqrt(n)



* To figure out t 🡪 need to determine the dF associated w/ the t-distribution needed for this data
* When working w/ data *from only 1 sample* + *estimating a single mean*, the dF = n-1.
* We lose 1 dF b/c we're *estimating the SE* of the sample mean *using the sample SD*.
* Putting all of this together, the CI for a *single population mean* can be estimated using x.bar +/- t\* w/ n-1 dF \* s /Sqrt(n)



* There are variety of ways of finding the critical t score 🡪 t-table w/ dF = 22 – 1 = 21 for the row + corresponding tail area for desired confidence level.

* If we a 5% confidence level, we have 95% of our data in the center of the distribution (want the middle 95% in our CI), so we have 5% left for the 2 tails.

> ## find critical value of t for sample size of 22

> n <- 22

> dF <- n - 1

> abs(qt(p = .025, df = dF)) # find percentile

[1] 2.079614

* Note we *always use a positive critical value* + the confidence level = always the middle symmetric area in the center of the curve
* Once you mark that, you can easily determine the tail areas + use that value to find critical t-values We finally have all of our building blocks
* Now construct the CI for the average snacking level of distracted eaters.

> x.bar.s <- 52.1

> s <- 45.1

> t.crit <- abs(qt(p = .025, df = dF))

> SE <- s / sqrt(n)

> mOe <- t.crit \* SE

> (lower <- x.bar.s - mOe)

[1] 32.10378

> (upper <- x.bar.s + mOe)

[1] 72.09622

* **We are 95% confident distracted eaters consume between 32.1 to 72.1 grams of snacks post meal.**
* Next, suppose suggested serving of these biscuits = 30 g.
* *Do these data provide convincing evidence the amount of snacks consumed by distracted eaters post lunch is different than the suggested serving size?*
* Givens = sample mean, sample SD, sample size, SE calculated earlier = 9.62.
* Null: the population mean mu = 30 grams
* Alternative: mu != 30 (interested in any difference from mu, i.e. in either direction)
* The test statistic, t, can be calculated as sample mean - the null value divided by SE

> null.mu <- 30

> (t <- (x.bar.s - null.mu) / SE)

[1] 2.298408

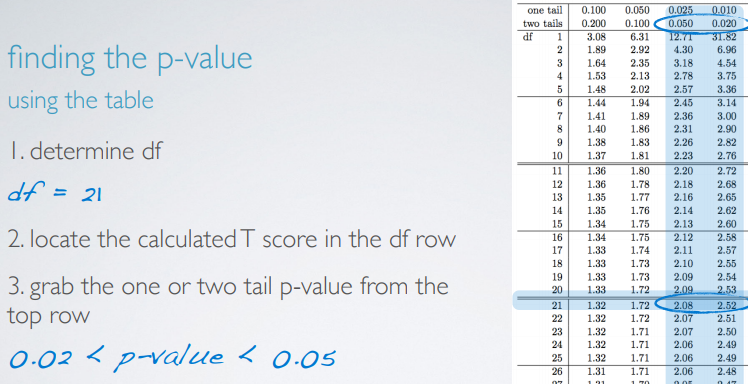
* Our observed test statistic = 2.3 *and* -2.3 (2-sided alternative hypothesis = shade both tails)

> # probability of obtaining this mean x.bar.s t w/ 21 dF if null = 30 is true

> pt(t, dF, lower.tail = F)\*2

[1] 0.03190849

* Or w/ table



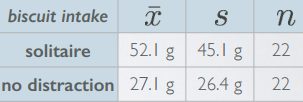
* Focusing on row of the table for our dF, locate the calculated t-score = 2.3 (work w/ the absolute value of the calculated t-score) + we grab the 1 or 2-tailed p-value from top of the table (depending on our alternative).
* In this case, we had a 2-sided alternative so our p is going to be somewhere between 0.02-0.05.
* This answer is less precise than the exact value R gives, but we still have sufficient info on the p-value to compare it to the significance levels of the test + make a decision.
* To recap, we focused on 1 group from the study (distracted eaters) + were provided some sample statistics on this group
* We calculated a 95% CI ranging from 32.1 to 72.1 g + did the hypothesis test where we compared how much these people ate to the suggested serving size.
* We found a p-value = ~3.18%, which < standard significance level of 5% = rejected the null + concluded these data DO provide convincing evidence distracted eaters consume an amount different than the suggested serving size.
* Since both *the estimation + the testing* were done using the *same underlying inferential framework* + the *same distribution*, the results should agree w/ each other.
* The null sets mu = 30 + we rejected this null.
* Similarly, the CI does NOT contain the null value of 30.
* Therefore, these two methods agree.
* 1 important task we skipped over = initially checking the conditions.
* We DO have a random assignment + 22 < 10% of all distracted eaters (we can assume).
* Therefore, we assume that 1 distracted eater in the sample is independent of another
* We're not given a visualization of the distribution of biscuit consumption to check the sample size skew condition.
* However given the sample statistics, we can kind of sketch it out.



* Sample mean = 52 + there's a natural boundary at 0 (one cannot < 0 g of biscuit)
* The 68, 95, 99.7 rule is not going to apply here (> 1 SD below the mean = hits natural boundary of 0 g)
* Therefore, the data are likely right-skewed
* The t distribution is pretty robust of skewness, but ideally we’d like to see a visualization of this distribution + asses this sample size condition accordingly, especially given the low sample size.

**Inference for Comparing 2 Independent Means**

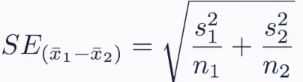
* Stick w/ the distracted eaters study.



* Want to estimate how much more/less distracted eaters snack compared to non-distracted ones
* Need a **CI** of this = Point estimate +/- Margin of error = (different between 2 sample averages) +/- critical value \* **standard error of difference between the 2 sample means**



* This is a new SE = square root of sum of squared variances divided by sample size of each group



* Note we *add* 2 variances even though we're looking for the SE of the *difference* of the 2 means
* Conceptually, think about it as bringing together 2 measures w/ an inherent variability around them = the 2 sample means.
* **When you bring 2 unknowns together, the result should always be *more variable*, regardless of whether you're adding them or subtracting them.**
* Also have a new dF calculation for 2 *INDEPENDENT* means



* This is actually not the *exact* dF, which is quite tedious to compute by hand.
* This is a conservative estimate, since it relies on the lower of the 2 sample sizes.
* Conditions to meet (like all inferential methods).
* 1) Independence, both within + between groups.
* Verified w/in-group independence *w/ respect to the outcome variable* via random sampling/assignment + the **10% condition** (if sampling w/out replacement)
* both n1 + n2 should be < 10% of their respective populations. I
* Failure to meet the between-groups independence condition is not inherently a problem
* just means that we’d need to use methods suited for *dependent* (**paired**) groups.
* We will introduce these
* 2) sample size and skew.
* More skewed the population distributions are = the larger the samples we need from them
* Estimate difference between the average post-meal snack consumption between those who eat with + w/out distractions.
* **CI = point estimate** (difference between the 2 sample means) **+/- a margin of error** (critical T-score \* SE of the difference between the 2 sample means.

> # find t-critical value for comparing the 2 means

> n1 <- 22

> n2 <- 22

> dF <- min(n1 - 1, n2 - 1)

> (t.crit <- abs(qt(p = .025, df = dF)))

**[1] 2.079614**

> # get point estimate

> x.bar.dist <- 52.1

> x.bar.nondistract <- 27.1

> (pe <- x.bar.dist - x.bar.nondistract)

[1] 25

>

> # get margin of error

> var.dist <- 45.1

> var.nondist <- 26.4

> (se.diff.2.means <- sqrt(((var.dist^2)/n.dist) + ((var.nondist^2)/n.nondistract)))

[1] 11.14159

> (m0e <- t.crit\*se.diff.2.means)

[1] 23.1702

>

> # get CI

> (lower <- pe - m0e)

[1] 1.829798

> (upper <- pe + m0e)

[1] 48.1702

* The **CI = {1.83, 48.17} grams.**
* Next, we need a hypothesis test for evaluating whether these data provide **convincing evidence** of a difference between average post-meal snack consumption between those who eat w/ + w/out distractions
* When doing a hypothesis test, 1st step always = **set your hypotheses.**
* H0 says there's absolutely nothing going on here + the difference between the average snack consumption for those who eat / + /out distraction = 0.



* Note we use μ + not x.bar b/c **hypotheses are always about populations + never about samples**
* We already *know the sample statistics*, so we *don't need to hypothesize about them*.
* Want to *use sample statistics* to *say something about the unknown population parameters.*
* The alternative = there *is* a difference between the 2 population means/is not 0.

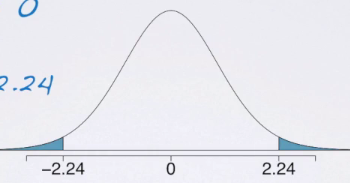


* Our T-score w/ 21 dF = the observed difference - null value of 0 / the SE

> (t.21 <- (pe - 0)/se.diff.2.means)

[1] 2.243845

* The last step before making a decision on these hypotheses = find the p-value



> ## probability of obtaining this OBSERVED DIFFERENCE (point estimate) w/ 21 dF if null = 0 is true

> pt(t.21, dF, lower.tail = F)\*2

[1] 0.03575082

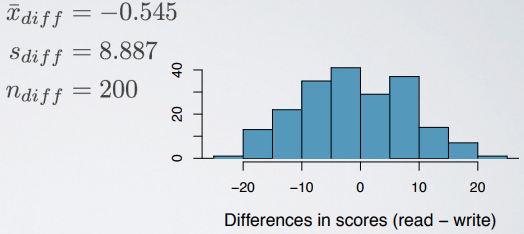
* To recap, we had / a study where researchers **randomly assigned** respondents into distracted + non-distracted eating groups + compared snack intake post-meal.
* The sample statistic suggested distracted eaters consumed more snacks on average.
* However, just b/c we observe a difference in sample means doesn't necessarily mean there is something going on that is **statistically significant** in the **actual populations**.
* So, we use **statistical inference tools** to evaluate if this apparent relationship between distracted eating + snacking more provide evidence of a **real difference** at **the population level**.
* Note, we have a **randomized control trial** here, so *if we do indeed find a significant result*, we could then talk about a **causal relationship** between these 2 variables.
* The **CI for the average difference** = {1.83, 48.17} + the hypothesis test evaluating a different between the 2 means yielded a p-value of roughly 4%.
* This means we’d reject the null (which is not in our CI) + conclude these data do indeed provide convincing evidence there is a difference between average snack intake of distracted + non-distracted eaters
* **The results of the CI + hypothesis test agree, as we used similar methods + we rejected the null (which set the difference between the 2 means to 0) + this null value was not included in our CI)**

**Inference for Comparing 2 Paired Means**

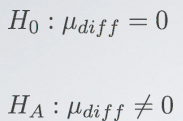
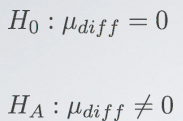
* Our methodologies change if the means we're comparing are **paired**/dependent (not much new)
* 200 observations were randomly sampled from the High School and Beyond survey.
* The same students took a reading + a writing test.



* At first glance, it appears median writing score is slightly higher than median reading score.
* Both distributions seem fairly symmetric, but reading score = slightly more right-skewed (evidenced by the fact that the median is closer to the 25th percentile than the 75th percentile)
* Reading scores are slightly more variable than the writing scores (wider box)
* *That all being said, at a first glance, it is really difficult to tell if there's a difference in the scores*
* So can the reading and writing scores for a given student be assumed to
* A student's reading score is likely NOT independent of their writing score.
* Generally high achieving students are likely to score highly on both tests.
* When 2 sets of observations have a special correspondence (are NOT independent) they are **paired**.
* To analyze paired data, it is often useful to look at the *difference in outcomes of each pair of observations.*
* Ex: For each student, subtract writing score from reading score + create a new variable = **diff** = the difference between the 2 scores 🡪 calculate this difference for each student in our data set.
* Good idea to start by defining the **parameter of interest** + the **point estimate**.
* In this case, we're interested in finding **average difference between reading + writing scores of all high school students = μ.diff**
* Since we don't have access to the whole population, we **estimate this unknown population mean w/ our sample statistic** = the average difference between reading + writing scores of these 200 sampled high school students, **x.bar.diff**
* If there was no difference between reading writing scores, we’d expect diff to be 0
* Look at the distribution of these differences.



* We see they’re centered around 0 but the average difference is not exactly equal to 0 + we're also seeing quite a bit of variability in this distribution.
* Therefore, it's impossible to determine whether there is a statistically significant difference between average reading + writing scores simply by visually evaluating this plot.
* We need statistical inference tools once again 🡪 1st define our hypotheses.
* H0: there's nothing going on
* H1: there’s a difference between the test scores.

* We summarized our 2 columns of data (scores) into just 1 column of differences 🡪 once again we’re setting out to do inference on a *single* population mean μ.diff.
* Therefore, the structure = exactly the same as hypothesis test for doing a test on any single mean, except **we're really doing inference on a difference of paired means.**
* The mechanics, the conditions, etc. = all the same as working w/ a single population mean.

> mu.diff <- 0

> x.bar.diff <- -.545

> s.diff <- 8.887

> n.diff <- 200

> (dF <- n.diff - 1)

[1] 199

> (se.diff <- s.diff/sqrt(n.diff))

[1] 0.6284058

> (t.199 <- (x.bar.diff - mu.diff)/se.diff)

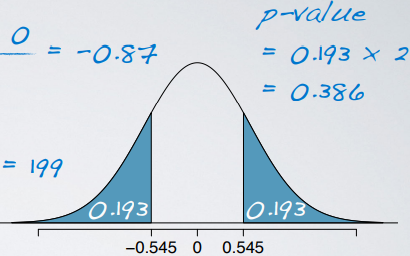
[1] -0.867274

* Then we draw our curve, mark the observed difference, + shade the tail areas corresponding to the p-value (2-sided alternative = shade both tails\_

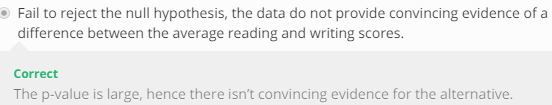
> ## probability of obtaining this OBSERVED DIFFERENCE (x.bar.diff) w/ 199 dF if null = 0 is true

> pt(abs(t.199), dF, lower.tail = F)\*2

[1] 0.3868365



* Compare the p-value to the significance level + if lower, reject the null + conclude the data DO provide convincing evidence for the alternative
* However, understanding what the p-value *actually means as a* ***conditional probability***, it usually takes a little bit more practice.
* 



* The best approach for finding the correct interpretation of the p-value as a probability are questions:
* ~~p-value = probability the average scores on the reading + writing exams are equal~~
* Basically means p-value equals the probability of the null being true = FALSE
* ~~p-value = probability the average scores on the reading + writing exams are different~~
* i.e. probability of the alternative being true = FALSE
* **p-value = probability of obtaining a random sample of 200 students where average difference between reading + writing scores is *at least* 0.545 in either direction if the true average difference between scores is in fact 0.**
* Generically reads as “probability of observed or more extreme outcome given the null is true” 🡪 indeed the definition of the p-value.
* ~~p-value = probability of incorrectly rejecting the null if in fact the null is true.~~
* This is actually the **probability of a type I error** + not the definition of p-value.
* In summary, we started off with 2 variables (reading + writing scores of the same set of students) + summarized these variables into 1 by taking the **pairwise differences**.
* In situations where we do inference for paired data, most often the null sets the average difference between the 2 paired means to = 0 (indicating no difference between them)
* Paired data can happen when we have a set of data from the same set of people, like in pre-post studies, such as a weight-loss study (post-weight will necessarily be dependent on pre-weight)
* Other studies might also take **repeated measures** on the same set of people, such as reaction time of the same set of people after they have spent the recommended amount of 7.5 hours the previous night or if they've only spent 2 hours.
* We might also use paired approaches when *we have different sets of subjects to begin with, but for some reason we believe these subjects to be not independent 🡪* **Twin studies** is an obvious example for these, or studies on partner A + partner B who are in a relationship.
* We’d design these studies as paired if we believe these individuals in the 2 groups are similar on *certain aspects* + we're evaluating their differences on other aspects.

**Power**

* Oftentimes, in experiment planning, there are 2 competing considerations:
* We want to collect enough data so that we can *detect important effects*
* But collecting data can be expensive + experiments w/ people may have some risk to patients.
* **Clinical trial =** health-related experiments where subjects = people + we work on determining an appropriate sample size where we can be 80% sure we’d detect any important effects of the drug.
* In other words: 🡪 **find the required sample size that will result in a test with 80% power**.
* 80% might seem arbitrary, but it is indeed a commonly required power for most experiments.
* When we make a decision on a hypothesis test, 1 of 4 things can happen.
* Null is rejected when actually true = **Type 1 error** = **FP**.
* probability of a Type 1 error = the **significance level** of the test, alpha **α**.
* set at the beginning of the test.
* Null is failed to be rejected when indeed true = Right decision is made
* Probability of this = complement of the significance level = 1 – α
* Null is failed to be rejected but the alternative is actually true = **Type 2 error** = **FN**
* Probability of a Type 2 error = beta **β** (more complicated to calculate)
* Null is *correctly* rejected
* Probability of this = the **power** of the test = complement of Type 2 error rate = 1 - β
* Therefore, **keeping Type 2 error rate low increases the power**, a desirable outcome.
* In a hypothesis test 🡪 want to keep error rates low, BOTH α + β.
* However, *decreasing one increases the other*
* 1 solution for this = getting a larger sample size.
* Hence, *it's important to think about sample size when designing an experiment +* making sure resources are invested to recruit sufficiently large # of subjects to obtain the desired power of a test
* Suppose a pharmaceutical company has developed a new drug for lowering BP + are preparing a clinical trial to test the drug's effectiveness.
* They recruit people taking a particular standard BP medication
* ½ of the subjects are given the *new* drug = the **treatment group**.
* Other ½ continued to take their meds in generic-looking pills to ensure **blinding** = **control group**
* For this 2-sided hypothesis test:
* Null H0: There is no difference in average BP of treatment + control groups.
* Alternative H1: There is indeed a difference in average BP of treatment + control groups.
* 2-sided alternative hypothesis tests are common in clinical trials 🡪 interested in finding out if a new drug is better/worse than existing treatments.
* Suppose researchers would like to run this clinical trial on patients w/ systolic BPs between 140-180 mm of mercury + suppose previously published studies suggest the SD of patients' BPs will be ~12 mm of mercury + the distribution of patients' BPs will be approximately symmetric.
* W/ 100 patients/group, what would be the approximate SE for difference in sample means of the treatment + control groups?

> n.tx <- 100

> n.control <- 100

> dF <- min(n.tx - 1, n.control - 1)

> (t.crit <- abs(qt(p = .025, df = dF)))

[1] 1.984217

> # get margin of error

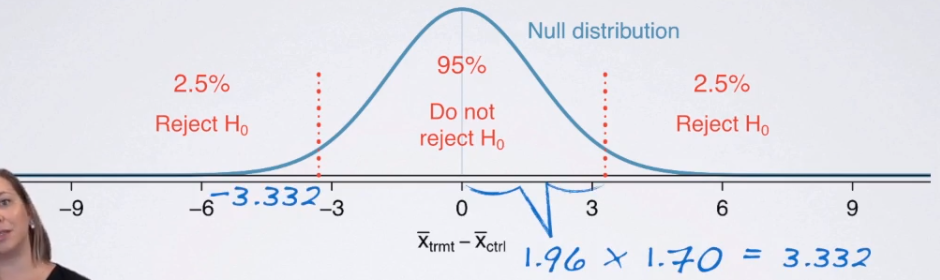
> var.tx <- 12

> var.control <- 12

> (se.diff.meds <- sqrt(((var.tx^2)/n.tx) + ((var.control^2)/n.control)))

[1] 1.697056

* In a test for comparing 2 ***independent*** means 🡪 calculate SE = sum of the variances of the 2 groups (SD-squared) divided by their respective sample sizes 🡺 SE = 1.70 mm of mercury.
* Then, according to the CLT, the distribution of the differences in sample means will be nearly normal, w/ mean 0, our null value.
* Using this info, we can find out what values of the sample statistic we’d need to reject the null

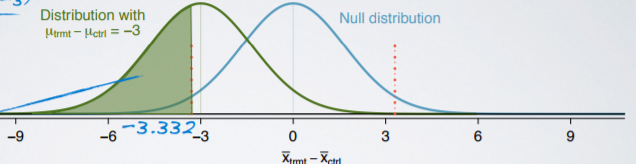


* 1st draw our null distribution = nearly normal + centered at 0 (the null value)
* Rejecting the null requires having a sample statistic sufficiently far from the null value such that the two tail areas will be less than 5% total.
* i.e. sample statistic needs to fall in the rejection regions
* Under the normal model, 95% of the observations fall w/in 1.96 SDs of the mean + since we measure variability of *this* distribution by the SE, the rejection region starts 1.96\*1.70 (critical value \* SE) = 3.332 mm of mercury away from the mean/null value

> (m0e <- t.crit\*se.diff.meds) # value to be away from mean to reject null

[1] 3.367328

* could be on the positive or negative side of the null b/c it’s a 2-sided alternative test.
* Suppose the company researchers care about finding ANY effect on BP that is 3+ mm of mercury vs. the standard medication.
* **What is the power of the test that can detect this effect** (the 3+ mm)**?**
* In other words, 3 mm of mercury = the MINIMUM effect size of interest
* We want to know how likely we are to detect this size of an effect in this study.
* If the treatment is *indeed effective enough* to result in a 3 mm of mercury drop in BP on average, it means the observed distribution of differences in average BPs between the 2 groups will be shifted from the null by 3 mm of mercury,



* We also know **we can only reject the null if the observed difference is < -3.332 mm of mercury.**
* Putting all of these together, probability of being able to reject the null if the true effect size is -3 is equal to the green shaded AUC above
* We've simplified calculating power to just calculating an AUC for the normal curve
* Calculate a Z score = difference in sample means needed (-3.332 - the mean of Tx distribution = -3) divided by SE

> min.effect.needed <- -3

> (z <- (-m0e - min.effect.needed)/se.diff.meds) # negative m0E b/c concerned w/ drop in mercury

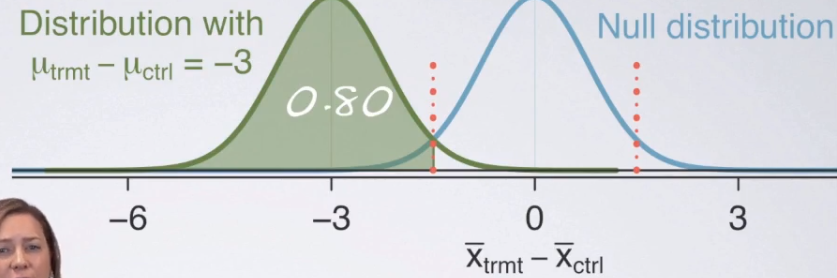
**[1] -0.21645**

> # find AUC = probability of being able to reject the null if the true effect size is -3

> pnorm(round(z,1))

**[1] 0.4207403**

* Therefore, the **power of the test** = ~42% when effect size = -3 + each group has a sample size of 100.
* Obviously, this is much lower than the 80% power we set out to attain
* It highlights how important it is to not just arbitrarily select a sample size + risk being left w/ an under-powered study.
* To fix 🡪 *work backwards from desired power to determine minimum required sample size*
* Note\*\*\* = Effect size is still = -3, since that's what the company is interested in
* However, SE will now be different since it changes when sample size changes.



* See we marked desired power on the green shaded area.
* Working backwards = 1st determine Z-score that marks 80th percentile of the normal curve.

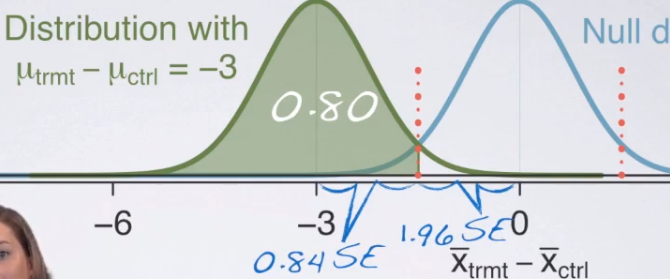
> # calculate z-score for desired AUC/power of .8

> desired.pwr <- .8

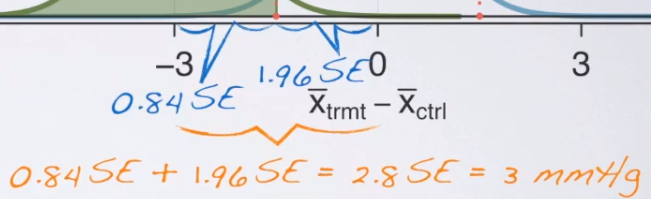
> (desired.z <- qnorm(desired.pwr))

**[1] 0.8416212**

* The 80th percentile is marked by Z = 0.84, therefore the distance between the center of the green distribution + the cutoff for the rejection region = 0.84 \* the SE of this distribution (still unknown)
* We know the distance between the center of the null distribution + the rejection region is 1.96 \* SE for a hypothesis test w/ 5% significance.



* ***Note***: We're assuming the SEs of the null distribution + of the distribution of the observed data are the same (this would be true if the drug only *lowers* BP but doesn't change its variability)
* So, **effect size of 3 mm of mercury is = 0.84\*1.96 = 2.8 SEs**



* Solve for 1 unknown 🡪1st calculate SE

> (effect.size <- desired.z + crit)

**[1] 2.825838**

> (new.se <- abs(min.effect.needed) / round(effect.size,1))

**[1] 1.071429**

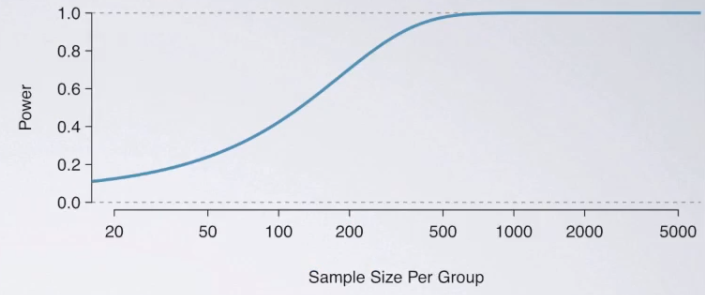
* Then set this value = the sum of the variances of the 2 groups (12 squared) / unknown sample sizes, which are the same

> # solve for new sample size via SE formula

> (new.n <- (var.tx^2 + var.control^2)/(new.se^2))

**[1] 250.88**

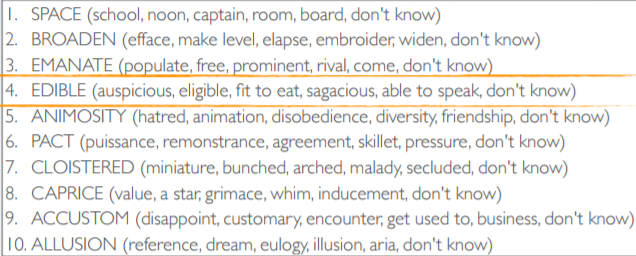
* We need at least 251 observations in each group in order to detect effect size of 3 mm of mercury
* *When are these calculations actually used in practice? 🡪* when designing a study to calculate a required sample size n for a desired level of power.
* Or can calculate power for a *range of sample sizes +* choose the target level of the power based on resources available for collecting the required sample size.



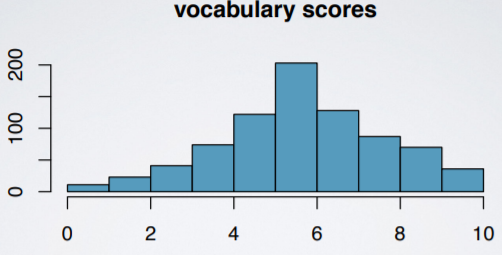
* This shows the power of the test we've been working w/ calculated for a sample sizes of 20-5k patients/group.
* Each data point on this curve = the power of the test for a given sample size
* As sample size increases so does power but only up to a point, + there seems to be no good reason to recruit > 500 patients or so for each group, since power plateaus at that point.
* This is important to know when designing a study in order to avoid wasting resources on a sample size larger than needed for maximum power desired.

**Comparing > 2 Means**

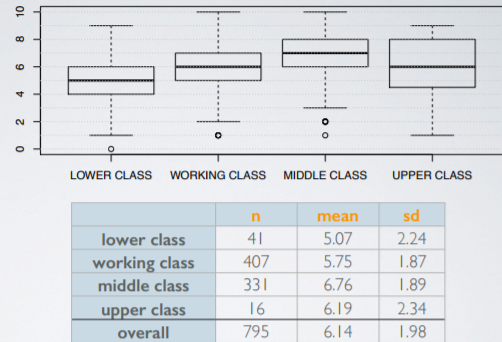
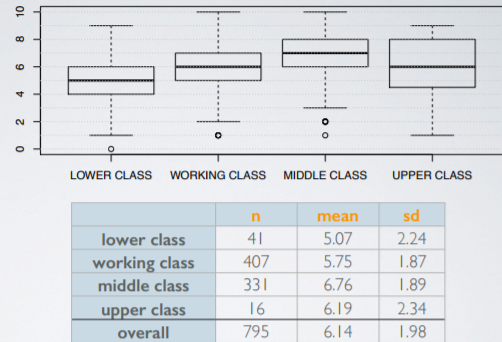
* General Social Survey w🡪 2 variables of interest = vocab scores + self-identified social class.
* Vocab score: calculated based on set of 10 question vocab test, where higher score = better vocab
* Respondents given a list of words + are asked to choose a word from the list that comes closest to the meaning of 1st word (capital letters).



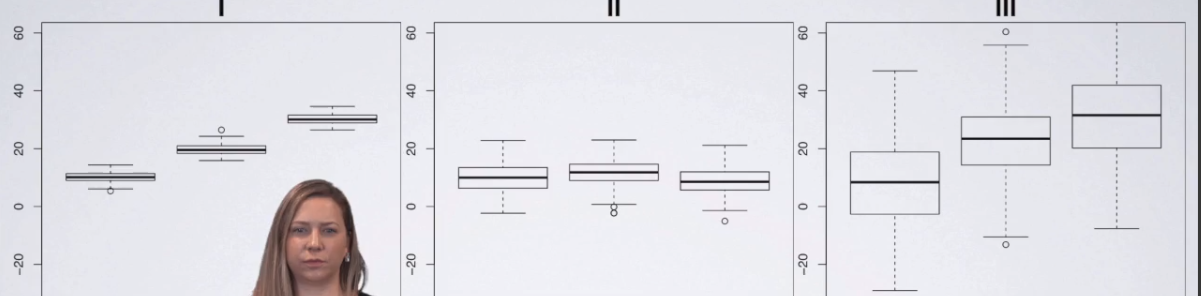
* Self-identified social class: 4 levels = lower, working, middle, upper class.
* We're focused on how people who took the survey did on the vocab test + whether scores are associated w/ social clause.

* These graphs tell us about variables *individually* but don't tell us much about their *relationship*.
* **Side-by-side box plots** = useful for visualizing relationship between numerical + categorical variable, + so are **summary statistics**

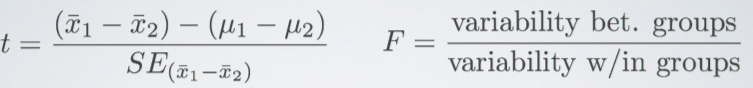
* We can see *some* differences between groups, but don't yet have the tools for determining whether these differences are statistically significant.
* Let's take a quick look at this question: Which of the following plots shows groups w/ means that are most + least likely to be significantly different from each other?



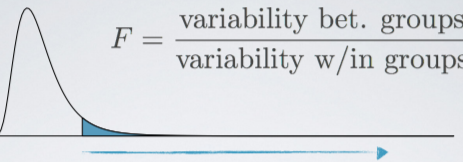
* Groups clearly separated from each other = most likely to have significantly different means
* Plots 1 + 3 = groups w/ same centers but **data in plot 1 is much less variable = much easier to detect differences in means for data in plot 1** b/c groups are much obviously separated.
* Middle = least likely to be significantly different = groups w/ very close centers
* Goal = find out if there's a difference between average vocab scores of different US working classes.
* Can compare means of 2 groups w/ t tests but comparing 3+ groups requires a new test = **analysis of variance** + a new statistic, the **F statistic**.
* **Null H0 in ANOVA** = just like any other null = “there's nothing going on” = **the mean outcome is the same across all categories.**



* k = # of groups = levels of explanatory categorical variable.
* Alternative H1 = “there is something going on” 🡪 not very specific.
* Says *at least 1 pair* of means are different from each other but doesn't specify which.
* For now, if we DO reject the null = we’ve found out there IS something interesting going on in the data + we might need to dig deeper to find out *which* group means are different from each other.
* **t-test = compare means from 2 groups to see if they're so far apart that the observed difference cannot reasonably be attributed to sampling variability**.
* **ANOVA = compare many means from > 2 groups to see whether they're so far apart observed differences cannot all reasonably be attributed to sampling variability**
* **t statistic** = **ratio** of **effect size** to **standard error**.
* ANOVA’s **F-statistic** = **ratio** calculated a bit differently b/c there isn't a single population parameter or point estimate we can ID b/c we're comparing many means.
* **F-statistic** = **ratio** of **variability between groups** to **variability within groups**.



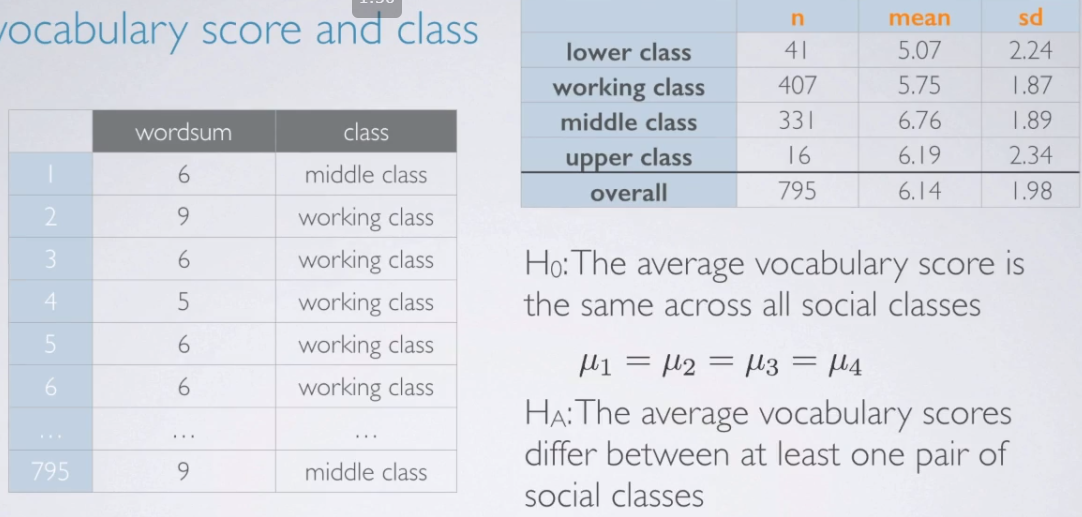
* Remember: Large test statistics = small p-values.
* As test statistics get closer to tails, the tail areas from the test statistic to the end of the curve get smaller + smaller.
* Remember = If p is small enough, we can reject the null + conclude the data provide convincing evidence for a difference in population means.
* **F distribution =** right-skewed + *always positive* (b/c it's a ratio of 2 measures of variability, *which can never be negative*)

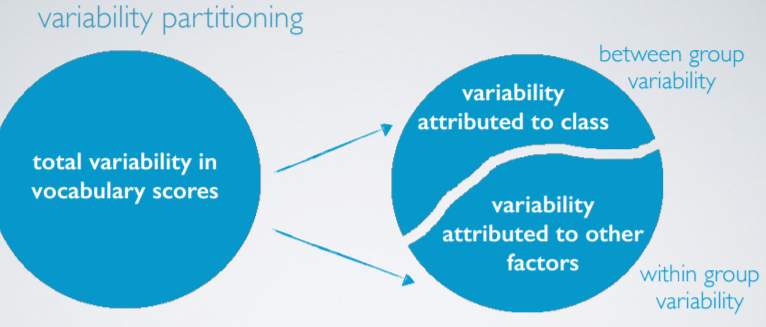


* **In order to be able to reject the null, we need a small p value which requires a LARGE F statistic**.
* Obtaining a large F statistic requires variability between groups be much larger than variability within groups.

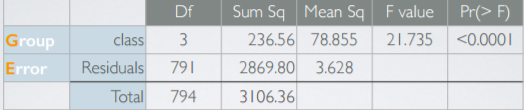
**ANOVA**

* **Variability partitioning** = means considering different factors that contribute to variability in our response variable.
* Ex: Variability in final exam scores will likely be due to a variety of factors.
* 1 might be if a student completed all components of a course leading up to the final,
* There will certainly be other factors as well = Familiarity w/ material beforehand, # of hours per week put into the course, so on + so forth.
* Suppose we're interested in studying how strongly completing all components leading up to the final is associated w/ exam scores.
* To study this, we’d **partition** the total variability in exam scores as variability due to this variable (completed all components) + variability due to all other factors.
* Remind ourselves of the data we're working w/.

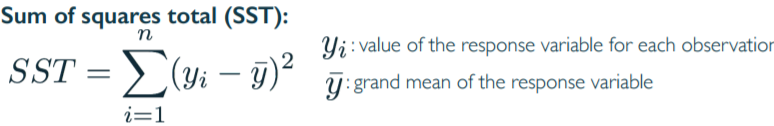




* So, we have total variability in vocab scores that we **partition** into 2 variabilities: variability in social class + variability attributed to all other factors.
* Variability attributed to social class = **between group variability** b/c social class = the **grouping variable**
* The other portion of the variability is what we're NOT interested in = **within group variability**
* Somewhat of a nuisance factor b/c if everyone w/in a social certain class scored the same, then we’d have no variability attributed to other factors = no w/in group variability
* ANOVA output table



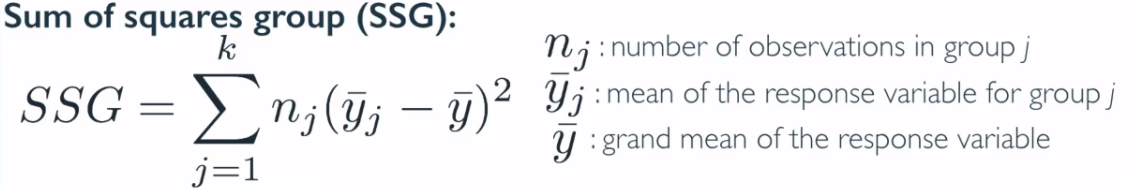
* 1st row = about between group variability = **group row**
* 2nd row = about within group variability = **error row**
* 3rd row = the totals.
* **Sum of squares total (SST) = total variability in response variable (vocab score)**
* Calculated very similarly to variance except that it is NOT scaled by sample size.
* = square deviation from the mean of the response variable.

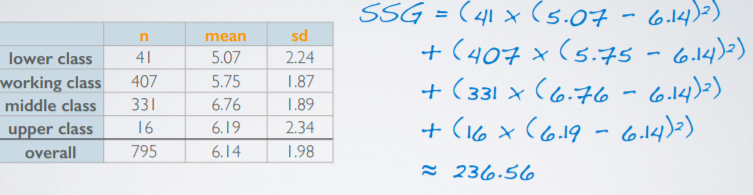


* We have 795 observations in our dataset w/ mean vocab score = 6.14
* To calculate SST, take each individual outcome score + subtract outcome mean = 6.14 from it, square that difference, + finally add up all those values.

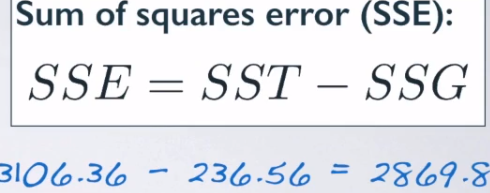


* This value, **3106.36 = total variability in the response variable**.
* What we're REALLY interested = how this *total* variability is *partitioned* into between + within group variabilities
* **Sum of squares group (SSG)** measures variability between groups = variability in the response variable that is *explained by explanatory variable in the analysis*.
* SSG = deviation of group means from the *overall mean* **weighted by their sample sizes**.
* i.e. For each group, calculate it's mean, **y.bar.j**, + substract the **grand mean** from it, **y.bar**, square this value, + multiply it for the sample size for that group
* Do this for each group + sum them up.

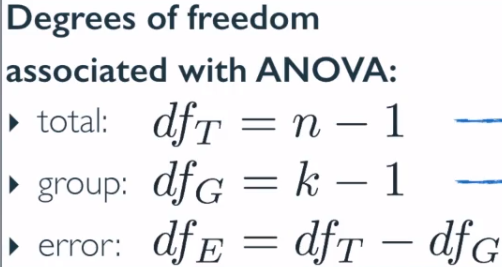




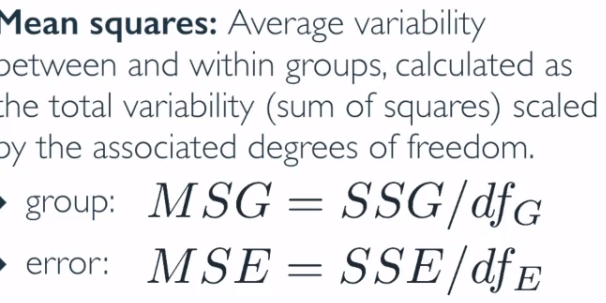
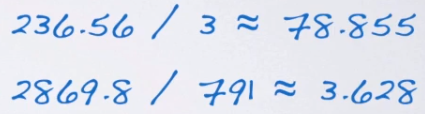
* SSG = 230.56 is not a meaningful # on its own but it's interesting to note how it compares to SST calculated earlier.
* This value 230.56 is roughly 7.6% of SST, 3106.36 **🡪 7.6% of variability in the outcome (vocab scores) is explained by the explanatory variable (social class) + the remainder is NOT explained by the explanatory variable (social class)**
* *= total variability explained by the groups*
* This is a low %, which makes sense b/c we’d expect vocab scores to be associated more so w/ education or how much people read than w/ social class
* **Sum of squares Error (SSE)** = variability within groups = the *unexplained* variability due to all other variables.
* Simplest way of calculating this = difference between SST + SSG.



* Now we need a way to get from the SS values to the **mean square (MS)** values
* Do this by *scaling* SS values by values that incorporate sample size n *as well as* # of groups (namely the dF)
* Total dF = n – 1 = **794**.
* Group dF = # of groups - 1 = **3**
* Error dF = the difference between the above = **791**.



* **Mean squares (MS)** = *average* variability between + within groups
* = SS for that component divided by dF

* Use MS values for calculating our **F score** = ratio of the average between + within group variabilities
* **F = MSG / MSE.**

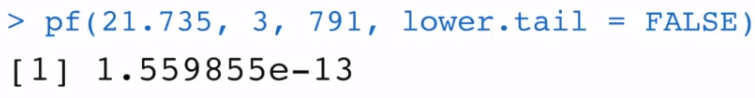


* Once you have your F score = ready to find p-value + conclude the hypothesis test.

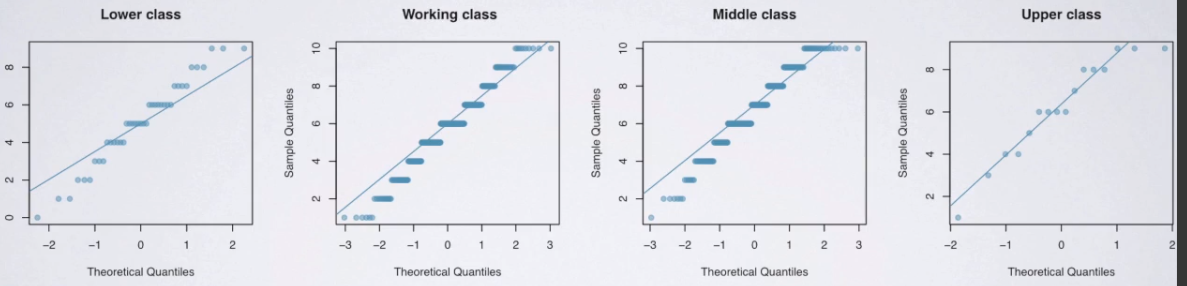


* The **p-value in this context = probability of *at least as large a ratio* between the between + within group variabilities if in fact the means of all groups are equal.**
* p-value = probability of observed or more extreme outcomes given the null is true.
* p-value = area under the F distribution.
* F statistic has 2 dF 🡪 group + error
* So, the p-value for an ANOVA = tail area under the F distribution W/ 3 + 791 dF = *tiny*.
* Even though we're looking for differences, we **only consider the upward tail of the F distribution b/c the F statistic can never be negative**.
* The ratio of 2 measures of variability can't ever be negative either.
* Since the F statistic is always positive, **a more extreme statistic will always be more extreme in the positive direction.**

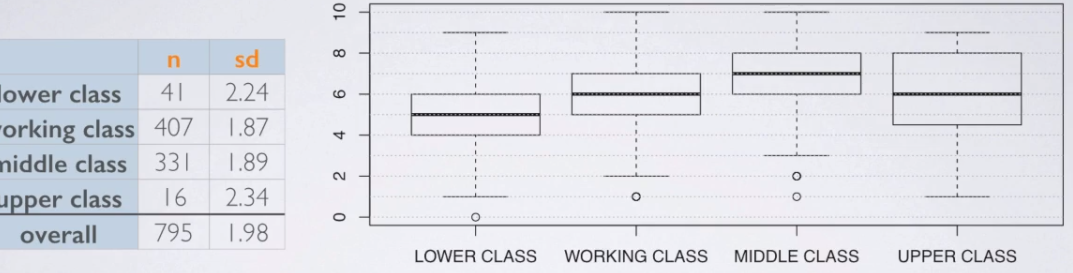




* If the p-value is small, we reject the null + say we have sufficient evidence for the alternative.
* If the p-value is large, we fail to reject the null + conclude the data do not provide convincing evidence at least 1 pair of population means are different from each other + that the observed differences in sample means are then attributable to sampling variability or chance.
* 3 main conditions for ANOVA
* Independence.
* Within groups 🡪 sampled observations w/in a group must be independent of each other
* Can assume this to be the case w/ random sample (or assignment, depending whether an observational study or an experiment) + if each sample size is < 10% of its respective population (if sampled without replacement)
* Sometimes can be difficult to check if we don't have sufficient info on how study was designed + how data were collected
* Between groups 🡪 groups should be independent of each other (non-paired)
* Checking this requires careful consideration on whether there is a paired structure between groups.
* If yes, this is not the end of the world but requires a different + slightly more advanced version of ANOVA = **repeated measures ANOVA**.
* Approximate normality 🡪 distribution should be nearly normal w/in each group
* condition is especially important when sample sizes are small but it's also difficult to check when sample sizes are small.
* Can visually check this condition using **normal probability plots**.



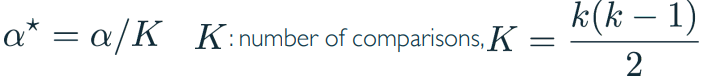
* See that lower + upper class plots are a little difficult to read b/c of lower sample sizes
* Middle class has quite a bit of divergence from normality in the lower tail 🡪 this condition may not necessarily be met.
* Homoscedasticity (equal variance) 🡪 variability of the distribution of the response variable w/in each group should have roughly equal variance.
* Constant variance across groups/variability = consistent across each group.
* Condition is especially important when sample sizes differ between groups.
* Side by side box plots + summary statistics are useful for checking this condition



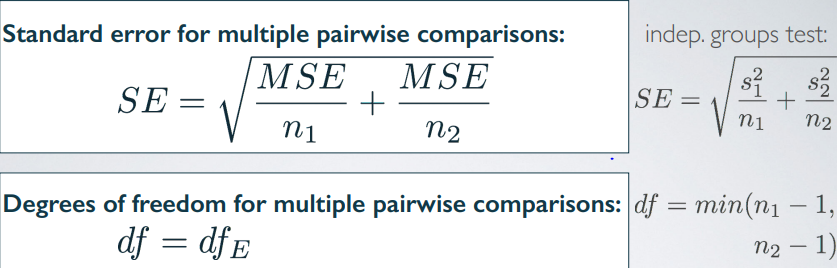
* Seems like the variability is consistent across lower, working + middle classes, but is much higher for upper class

**Multiple comparisons**

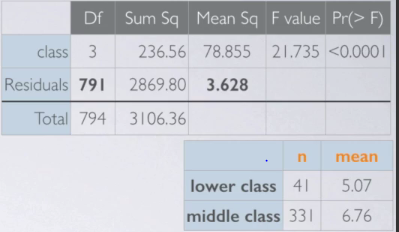
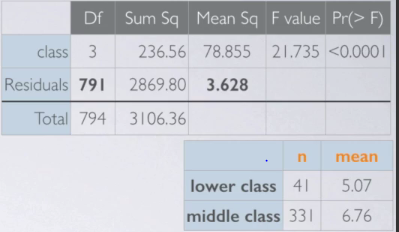
* ANOVA = method for comparing many means to each other concurrently.
* Finding a statistically significant result at the end of an ANOVA only tells us that *at least 1* pair of means are different, but not *which* pair of means are different.
* To determine whether 2 means are different from each other, we use t tests + w/ each test, you incur a probability of doing a **Type I (FP) error** (i.e. the significance level of the test, often 5%)
* So when doing multiple tests, you're inflating the Type I error rate, an undesirable outcome.
* Simple solution = Use a **modified significance level** lower than the original level for **pairwise tests** so that*overall* Type I error rate for the series of tests can still be held at the original low rate.
* Testing many pairs of groups = **multiple comparisons**
* Common modification = **Bonferroni correction =** a more stringent significance level for each pairwise test.
* Adjust **α** by dividing it by # of comparisons being considered.



* Ex: W/ 4 groups in an ANOVA that yields a significant result, you need to compare group 1 to group 2, group 2 to 3, group 3 to 4, so on + so forth.
* Counting these out = tedious + error prone, so we use a shortcut formula for determining this value + then use this value to adjust the significance level.
* Bonferroni correction = 2-step process.
* 1) Find the # of comparisons, K = (k – 1) / 2
* 2) Correct your original alpha by this level,
* Social class has 4 levels (k = 4) + if the original alpha was .05, what is going to be the modified significance level for the multiple comparisons test?
* Each group needs to be compared to 3 other groups 🡪 (4(4-1))/2 = 12 comparisons.
* But, if you have already compared group 1 to group 2 (middle to lower), you don't need to go back + compare lower to middle.
* So, total # of comparisons can be cut in half. Hence, this results in 6 total comparisons
* Use this # to correct our significance level down **to .05 / 6 = 0.008 = the significance we're going to use** for the pairwise comparisons (to see if 2 pairs of means are different from each other)
* There are a couple other considerations when doing these multiple comparisons after ANOVA.
* 1st, the **constant variance condition**
* Since for ANOVA, we need to meet this homoscedasticity condition, we need to now rethink the SE + the dF to be used in the multiple comparisons test (keep consistent for all tests)
* Now have a new modified significance level to compare p-values of these tests to in order to determine significance.



* Now for SE, instead of individual group variances, use **mean squared error** from ANOVA table
* Remember, **MSE = average w/in group variance**, so it’s still the same thing (individual group variances) but now, we have a consistent measure we can use for all of the tests
* If the constant variance condition is indeed satisfied, this value should be very close to group variances anyway
* Consistent dF = **dF error** from the ANOVA output, as opposed to the minimum sample size from the 2 groups we're comparing - 1.
* Is there a significant difference between average vocab scores between middle + lower class Americans?
* H0: There is no difference + the averages are equal, +
* H1: Averages are different

> n.low <- 41

> n.mid <- 331

> x.low <- 5.07

> x.mid <- 6.76

> dF.btwn <- 3

> dF.w <- 791

> ss.b <- 236.56

> ss.w <- 2869.8

> ms.b <- ss.b / dF.btwn

> ms.w <- ss.w / dF.w # ms.w = MS errors

> ms.e <- ms.w

> (f <- ms.b / ms.w)

[1] 21.73426

> pf(f, dF.btwn, dF.w, lower.tail = F)

[1] 1.561427e-13

> x.diff <- x.mid - x.low

> mu <- 0

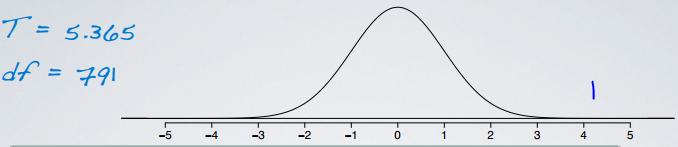
> (se <- sqrt((ms.e/n.low) + (ms.e/n.mid)))

[1] 0.3153575

> (t = (x.diff - mu)/se)

[1] 5.358998

* Write down test statistic + the dF (791) + before getting to p-value



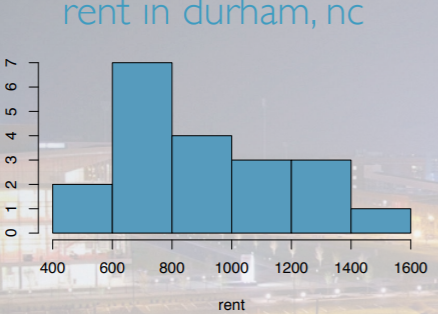
> pt(abs(t.791), dF.w, lower.tail = F)\*2

[1] 1.098564e-07

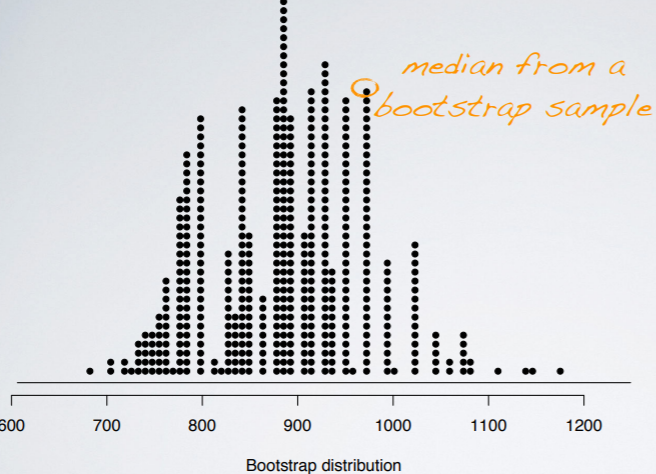
* Our T-score is very high (5 SDs away from center) + will result in a really, tiny, tail area b/c it's really unusual to get an observation > 5 SDs from the mean.
* This is also clear from the sketch w/ the really skinny tail areas when that far from the center
* Remember the significance level we used for this test is modified (.0008), + even though that was very conservative + stringent, b/c we got a pretty tiny p-value, we actually can reject the null again
* i.e. We can conclude the data provide convincing evidence that the average vocab scores of *self-identified* middle + lower class Americans are significantly different.

**Bootstrapping**

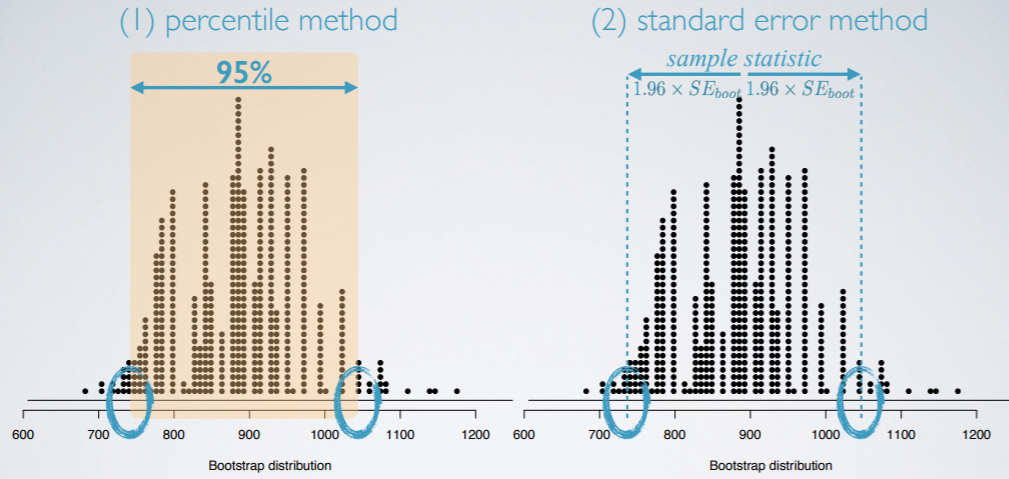
* **Bootstrapping** = a simulation-based method for inference, specifically for constructing CI’s
* We have a data set of 20 apartments randomly selected from Craigslist housing ads w/ at 1 bedroom in Durham, NC
* Look at the distribution of these rents,



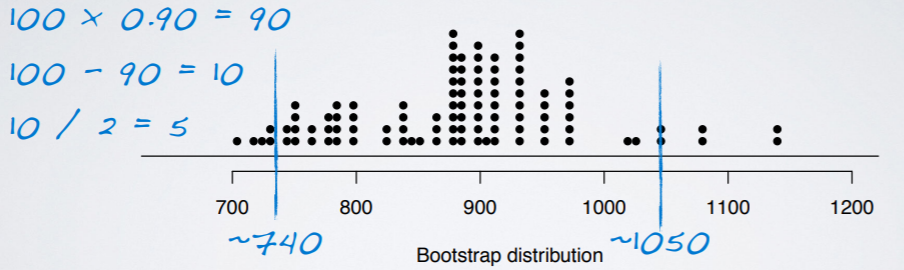
* Mean is probably a bit > $700, with median ~$700, and the median is a better measure of typical rent in Durham b/c the distribution of rents is right-skewed.
* The sample median is actually $887
* So, we cannot apply CLT-based methods learned so far to construct CI’s for the medium.
* A new method is indeed needed here = **bootstrapping** = a metaphor for accomplishing an impossible task w/out any outside help.
* In this case, the impossible task = estimating the population parameter (unknown population median) using *data from only the given sample.*
* This is something we've already been doing + it is what statistical inference is all about.
* We have a sample + we *use that sample* to make **inferences** about the *unknown population*.
* In bootstrapping, assume for each observation in the sample, there may be others like it in the population 🡪 think of bootstrap population as a population where each observation from the sample appears many times.
* Then, we take samples from *this* SIMULATED population to get an idea of how medians from the original population would look if we actually *could* take samples from it
* Don't actually create the bootstrap population but **simulate** it by taking **many samples from the original sample *with replacement***.
* Step 1: Take a bootstrap sample = A random sample taken **w/ replacement** from the original sample that’s the same size as the original sample.
* Step 2: Calculate the bootstrap statistic. = i.e. whatever statistic you're interested in.
* In the example, we were interested in the median, but it could be a mean, proportion, SD, etc.
* Step 3: Repeat these steps to create a bootstrap distribution = A distribution of bootstrap statistics
* This is just like creating the sampling distribution, but w/ 1 big difference 🡪 taking samples from the original sample instead of from the population, to which we don't have access.
* Here's our bootstrap distribution of medians constructed based on 500 simulations.



* Each dot represents a median from a bootstrap sample (remember each bootstrap sample is a sample of the same size as the original sample taken w/ replacement from the original sample)
* W/ replacement = it would be conceivable to get others like a specific observation in a sample if I had access to the original population.
* B/c we're sampling *w/ replacement*, each sample is going to be slightly different than the other, yielding a slightly different sample statistic.
* If we didn't do w/ replacement, we’d end up w/ the same sample + same median over + over
* Using the bootstrap distribution, we can calculate CI in 2 ways.
* 1) Percentile Method = estimate, say, a 95% CI, simply as the middle 95% of the bootstrap distribution
* Bounds of the interval = the 2.5th + 97.5th **percentiles** of the bootstrap distribution.
* 2) Standard Error Method = more accurate method 🡪 interval = the sample statistic +/- t\* by the SE of the bootstrap distribution **SE.boot**
* Critical T score will have n – 1 dF, where n = original sample size.



* See distribution of medians of 100 bootstrap samples from the original sample + estimate the 90% bootstrap CI for median rent, based on this bootstrap distribution, 1st using the percentile method
* Middle 90% of the distribution = the 90 observations in the center out of the 100 observations
* We only have 10 left for the tails = 5 on each side.
* Counting off dots from both sides, the CI comes out to be roughly $740-$1050.



* Look at the same distribution again + this time construct a CI using the SE method.

> n <- 100

> dF <- 19

> x.bar <- 887

> # 90% sig level

> (t.crit.90 <- abs(qt(p = .05, df = dF)))

[1] 1.792133

> se.boot <- 89.5758

> m0E <- t.crit.90\*se.boot

> (lower <- x.bar - m0E)

[1] 732.115

> (upper <- x.bar + m0E)

[1] 1041.888

* CI = ~$732-$1,042
* **We're 90% confident the actual median rent in the population of Durham apartments is somewhere in between these**.
* The percentile method + standard error methods yielded intervals pretty close to each other even though they're not exactly the same.
* **Bootstrap intervals** do NOT have as rigid conditions on sample size + skew as CLT-based methods.
* However, they yield more accurate estimates if sample size is larger (just like any other method)
* If the bootstrap distribution is extremely skewed or sparse, the bootstrap interval might be quite unreliable.
* Whatever method we use, simulation or CLT-bases, we still need a *good representative sample* from the population.
* **Simulation-based methods are NOT a cure for bad samples** 🡪 sample is bad = estimates will be bad
* Since bootstrapping sounds a lot like the process for building sampling distributions, compare + contrast:
* Sampling distributions = created using sampling w/ replacement *from the population*.
* Bootstrap distributions = created using sampling w/ replacement *from the sample*
* Both of these distributions = distributions of sample statistics.