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

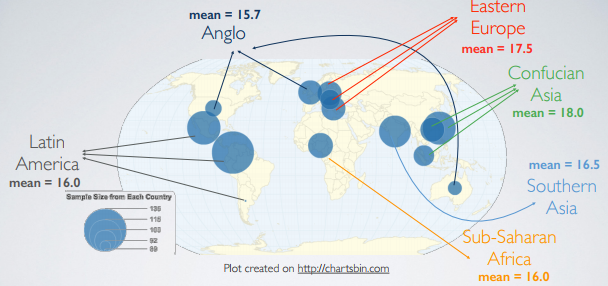
***COURSE 2 - Inference***

**WEEK 3 - Inference for Comparing Means**

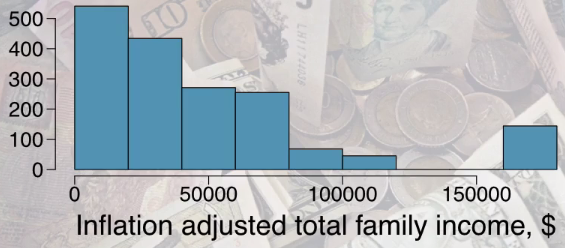
***4.3.1 t-distribution and Comparing 2 Means***

**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.



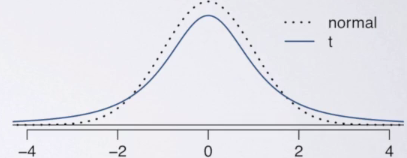
* See a geographic distribution of countries included in the study w/ sizes of circles = how large sample sizes (SS) are from each country.
* SS’s are somewhat consistent across globe + it seems like a pretty even geographic distribution
* Study further categorizes 14 countries into 6 continents + those 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



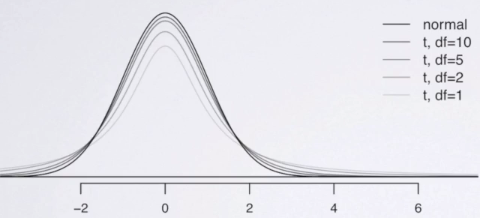
* Distribution is, as expected, pretty right-skewed.
* Suppose we‘d like to estimate typical total family income in the US.
* The CLT provided the basis for constructing a confidence interval for the mean, but what if we're not interested in the mean, ***but the median***?
* *No CLT for the median*.
* 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

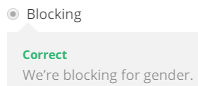
[1] 0.05094707

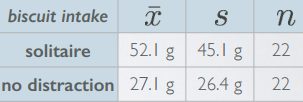
> pt(2, 10, lower.tail = F)\*2

[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 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



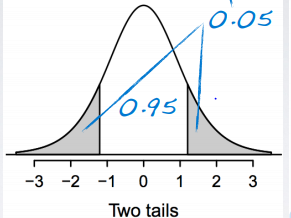
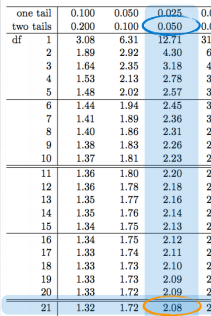
* This summary statistics suggest distracted eating groups snack more after lunch (x.bar.s = 52.1 g of biscuits compared to 27.1 g (x.bar.n) for the other group.
* We're also given the SDs for both groups, as well as the sample sizes, n, both = 22
* Goal = estimate average snacking level for distracted 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

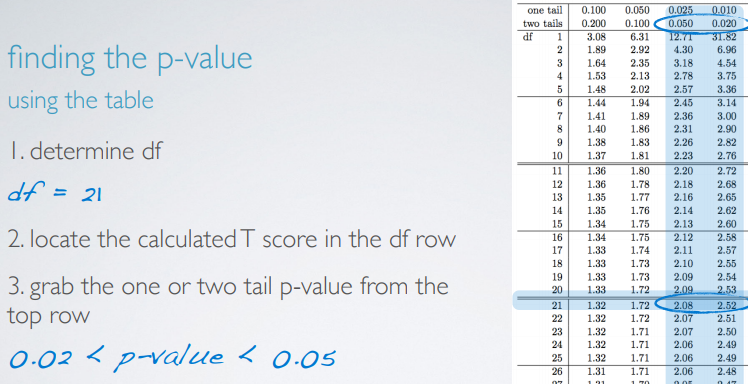
* Our observed test statistic = 2.3 *and* -2.3 (2-sided alternative hypothesis = shade both tails)

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

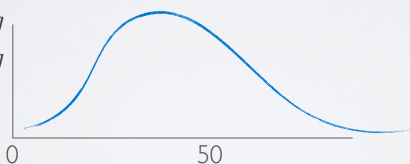
> pt(t, dF, lower.tail = F)\*2

[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.