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

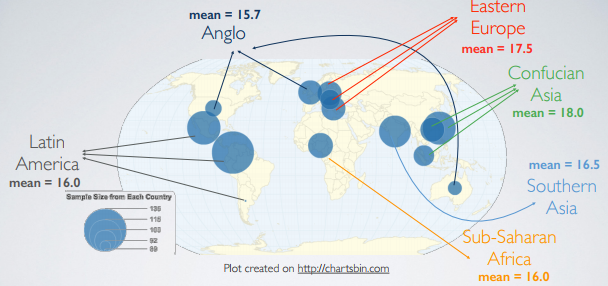
***COURSE 2 - Inference***

**WEEK 3 - Inference for Comparing Means**

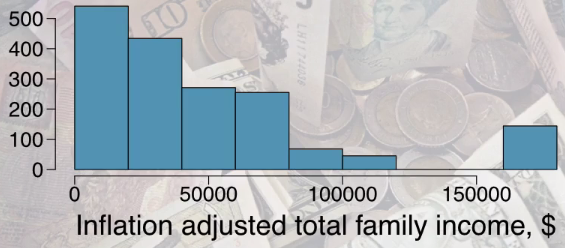
***4.3.1 t-distribution and Comparing 2 Means***

**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.



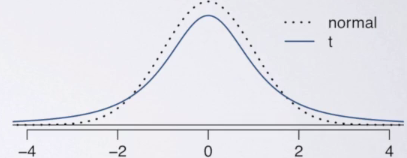
* See a geographic distribution of countries included in the study w/ sizes of circles = how large sample sizes (SS) are from each country.
* SS’s are somewhat consistent across globe + it seems like a pretty even geographic distribution
* Study further categorizes 14 countries into 6 continents + those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



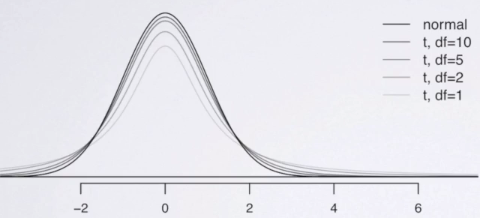
* Distribution is, as expected, pretty right-skewed.
* Suppose we‘d like to estimate typical total family income in the US.
* The CLT provided the basis for constructing a confidence interval for the mean, but what if we're not interested in the mean, ***but the median***?
* *No CLT for the median*.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