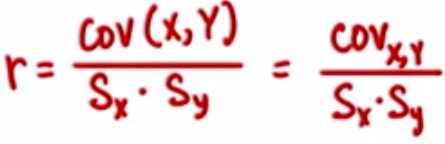
***Udacity Data Analyst Track***

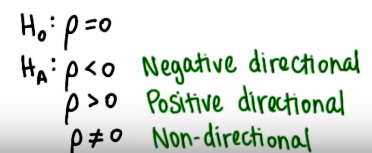
**I. Into to Inferential Stats**

9. Correlation

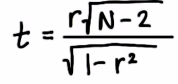
* So far, only been working w/ 1 variable, analyzing the mean + SD, and seeing where samples fall in the distribution + if that’s typical or not
* Now to focus on 2 variables which has a different design = need new to collect + visualize data, but analysis will be similar
* X variable(s) = **predictor, explanatory, independent variable**
* Y variable = **outcome, response, dependent variable**
* **Correlation Coefficient, Pearson’s r** 🡪 quantifies strength of relationship between 2 variables
* **r = covariance(x,y) / SD(x) + SD(y)**
* **numerator = covariance** = how much x and y co-vary (vary together)
* denominator = how much x and y vary apart from each other, rather than w/ each other



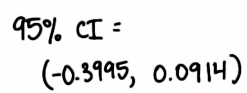
* Shown as ratio/proportion, not percentage
* So here, by increasing sample size, the size of r we need to reach significant gets smaller
* **Coefficient of determination, R2** 🡪 a percentage, specifically % of variation in y that can be explained by variation in x
* r measures strength of relationship by looking at how closely data falls along a straight line
* Perfect in positive direction = 1 Perfect in negative direction = -1
* No correlation 🡪 r = 0
* Ex: Age in years vs. Age in months 🡪 r = 1 b/c as one increases, as does the other in a straight line
* Ex: hours awake today vs. hours asleep today 🡪 r = -1 b/c as one increases, the other decreases a straight line
* Even if 2 variables appear to be related after looking at the data, it could just be due to chance
* The actual population might not have this relationship
* Or vice versa, where the population has a relationship but a sample from it does not
* Things like this happen due to sampling error, which can be reduced w/ larger sample sizes
* When doing t-tests, as sample size increased, we reached significance w/ a smaller difference between means b/c increasing sample size makes it harder to get a large difference by chance.
* But how much variation due to sampling error is allowed before we decide that results are NOT due to chance? 🡪 HYPOTHESIS TESTING
* r = sample data correlation
* **p(rho)** = true correlation for the population
* *ALWAYS FORM H(0) AND H(A) IN TERMS OF POPULATION PARAMETERS*, b/c that’s what we’re hypothesizing about



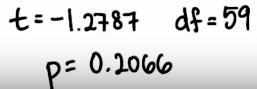
* Hypothesis testing about the *true* correlation of the population is a type of t-test
* t **= (r – Sqrt(N – 2))/ Sqrt(1 – r^2)** where dF = N -2 (b/c we have N values for each variable, + we subtract 1 per group)



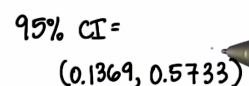
* Since we know sample statistics often aren’t exactly what pop. Parameters are, we might want a CI (range of likely values) for the true correlation (population coefficient, p(rho))



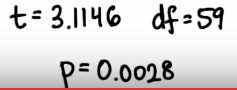
* This CI crosses 0 (contains 0 in the interval) 🡺 the true populatio correaltion could be 0, so there is not enough evidence to reject h(0) = p(rho) = 0



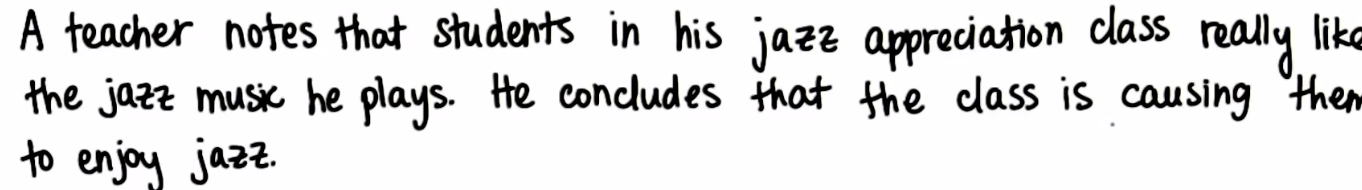
* Greater than alpha = 0.05



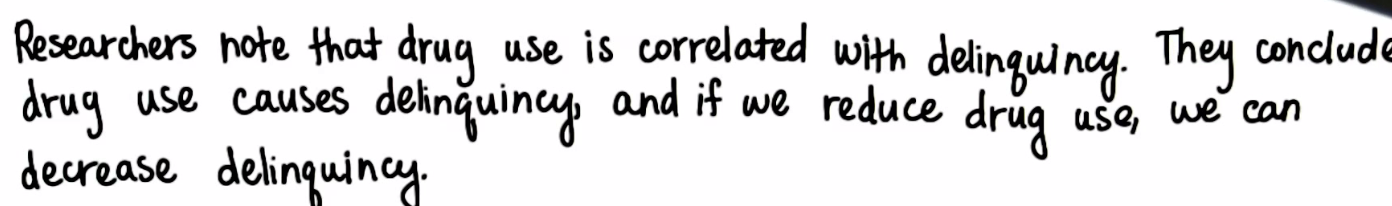
* 95% sure the true population coefficient is between these 2 numbers, and therefore not 0, so we have enough evidence to reject h(0)



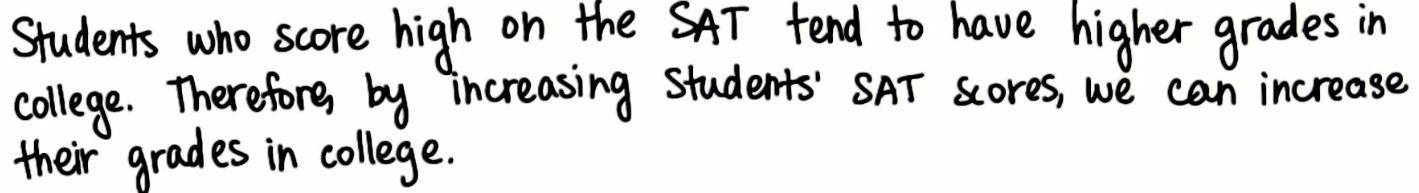
* less than alpha = 0.05
* ***REMEMBER 🡪 correlation does NOT equal causation***
* Variables x and y could both be influenced by another variable, A, **a confounding variable**
* Or variable x could influence y *through* variable A, **a mediating variable**
* To make CAUSAL statements the independent variable must happen *before* the dependent variable + we must be able to rule out any lurking variables



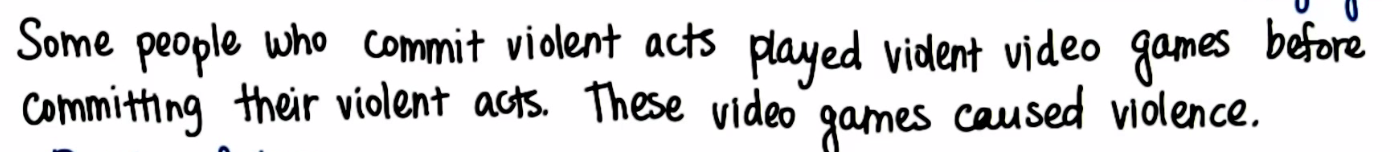
* Student could have liked jazz *before* taking the class (dependent happened before independent)



* ***Being delinquent could cause them to use drugs***
* The above 2 scenarios are examples of **ambiguous temporal precedence =** don’t know which variable occurred 1st



* **SAT happen before college, but there could be a third variable that influenced both SAT scores and college grades (motivation, intelligence, effort, etc.) 🡪 third variable problem**



* **Post-Hoc Fallacy** 🡪 just b/c someone played violent VG’s before violent acts, it doesn’t mean they caused it. People who want to commit violent acts also just want to play violent VG’s