Today's Agenda

Hour 1

Correlation vs association,

Pearson's R,

non-linearity,

Spearman rank correlation,

Hour 2

(Pearson) Correlation and regression.

Hypothesis testing for correlation and regression

Correlation and regression

Regression is used to further describe a linear relationship between two variables. The regression equation,

$$y = \alpha + \beta x + \epsilon$$

describes a model with three components:

α , The value that the y variable is when x is zero.

β , The amount that the y variable changes when x ***increases by 1***

ε , The error in the model.

$$y = \alpha + \beta x + \epsilon$$

The x variable is typically referred to as the ***explanatory*** variable, and the y variable as the ***response*** variable.*

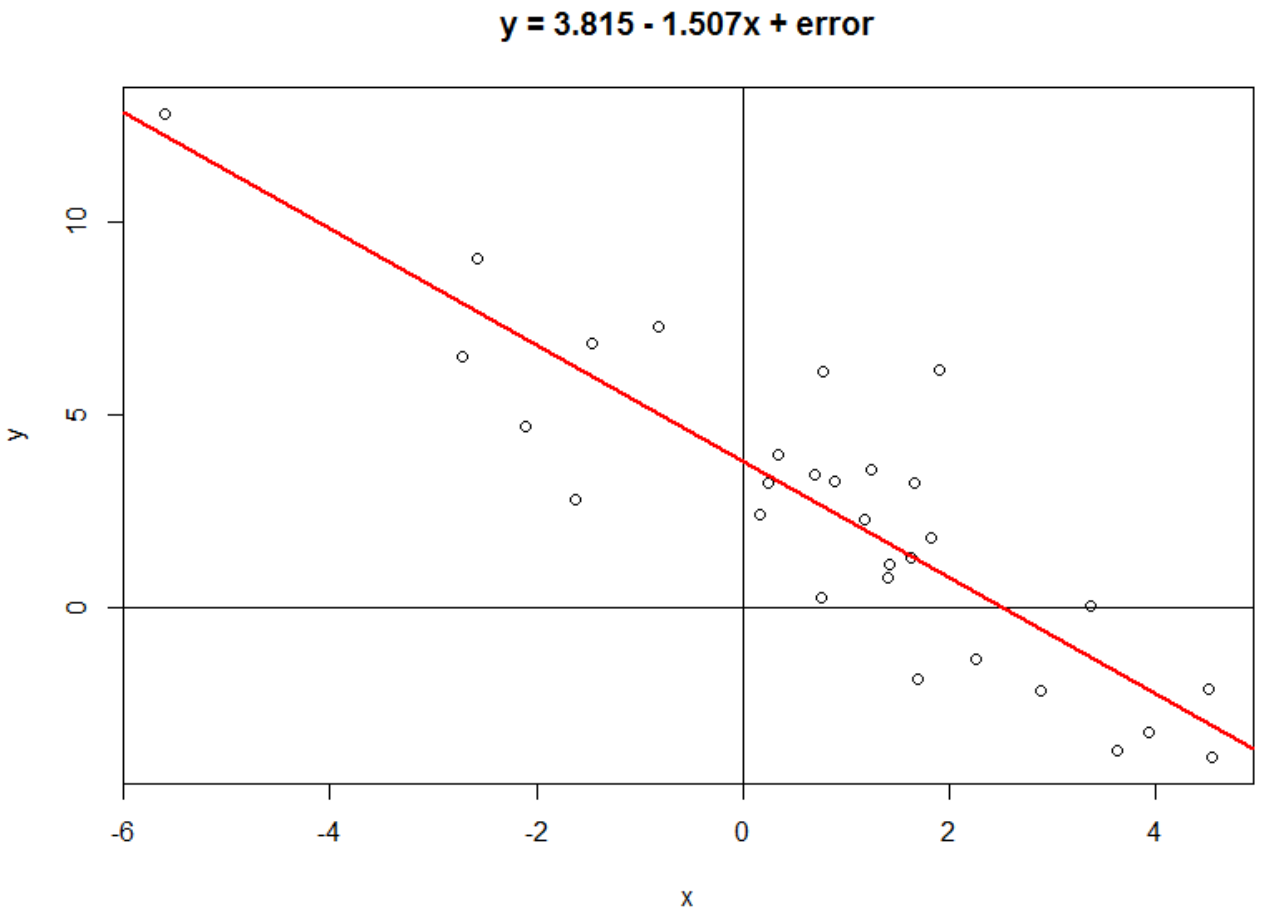
α and β are parameters, with estimates written as 'a' and 'b' respectively.

The error term ε (epsilon, pronounced epp-sill-awn), isn't a parameter because it's a different value for each observation. We will cover this a lot more later.

*See optional Reading 2.3: Variable names

The regression equation without the error term, α + βx ,

is called the ***line of best fit***, or the least squares line.



α is referred to as the ***intercept*** parameter. It represents that value that the line of best fit intercepts the y-axis. (i.e. where

x=0) In the above example, this value is estimated to be 3.815.

α always has a mathematical interpretation, but not always a concrete one.

The parameter α could represent the estimated heating costs of a home when it is 0° C outside, but it could also represent the fuel economy of a car with 0 mass, or 0 horsepower.

β is usually the parameter of interest, because if β = 0, then y does not (linearly) change when x changes. In the above example, this value is estimated to be -1.507.

β usually has a concrete interpretation.

It could represent the amount an average home's heating costs go up (down if negative) as the outdoor temperature increases by 1° C.

It could also represent the amount horsepower increases when the mass of a car increases by 1 kg. Or vice-versa.

Recall that Pearson correlation describes the strength and direction of a linear relationship. We can use the correlation coefficient to find a regression slope.

The slope,

$$\beta = \rho \frac{\sigma_y}{\sigma_x}$$

is estimated by

$$b = r \frac{s_y}{s_x}$$

It's all connected!

The formula linking correlation and regression is revealing...

$$\beta = \rho \frac{\sigma_y}{\sigma_x}$$

1. Slope is 'rise over run', which is the amount that y changes divided by the amount x changes.

This property is reflected in the ratio of standard deviations. The 'rise' sigma of y, divided by the 'run' sigma of x. If the data points were more vertically spread apart, the slope would increase.

$$\beta = \rho \frac{\sigma_y}{\sigma_x}$$

2. It is impossible to have less than zero variability in something. Therefore, measures of spread like σ, the standard deviation, are never negative.

Similarly, ratios of standard deviations, like σ_y / σ_x are never negative.

So, by process of elimination,

β is positive if and only if ***the correlation is positive***, and

β is negative if and only if the correlation is negative.

$$\beta = \rho \frac{\sigma_y}{\sigma_x}$$

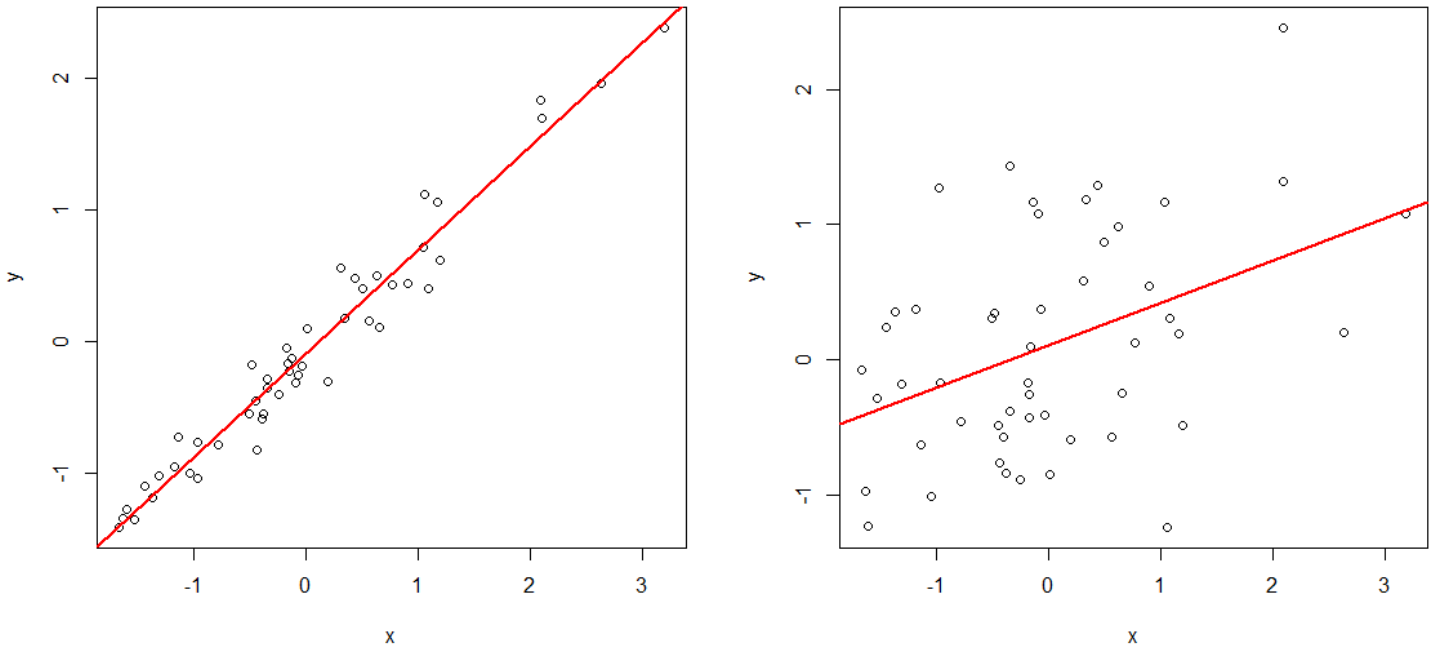
3. We assume that the response has some variance so,

$\sigma_y > 0$. ($\sigma = 0$ is an edge case, and not of concern)

Therefore, β can only be zero when the ***correlation is zero***.

2 and 3 In other words: The slope is the same sign as the correlation, and when there is no correlation x does not change (linearly) with y.

4. Finally, if the correlation is weak, such that ρ is close to zero, the slope is also close to zero. This is true even when y has a lot of variance.



The estimates of y regress toward to mean of x. In other words they are pulled back toward the mean.

Ready to regress?

Hypothesis testing for a correlation

or a regression slope.

Sometimes it's of interest to know if there is evidence of a correlation in your sample.

In others, you may want to test...

$H_o: \rho = 0$ against $H_A: \rho \neq 0$

Recall that if a correlation is positive/negative, then the associated regression slope is also positive/negative.

Likewise, if a correlation is ***significantly*** positive/negative, then so is the regression slope.

These two hypotheses are effectively the same:

$$\rho = 0 \leftrightarrow \beta = 0$$

...and they can be tested using the t-distribution.

This formula gives the t-score of correlation.

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

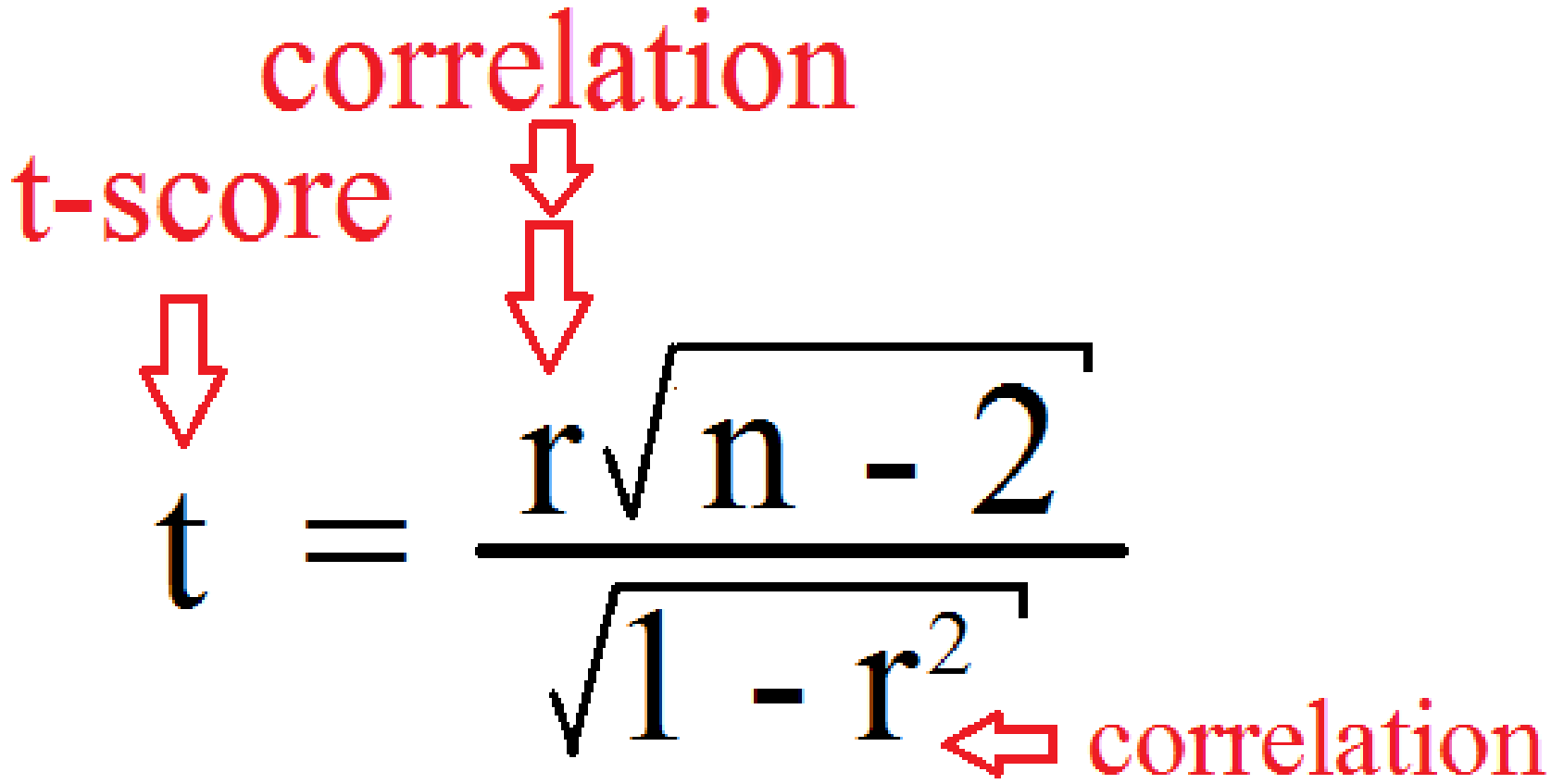
t-score

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

This t-score gets compared the critical values in the t-table at n-2 degrees of freedom. We use df = n-2 because we have n observations and 2 ***regression parameters*** to estimate.

The stronger the correlation, the farther r goes from zero.

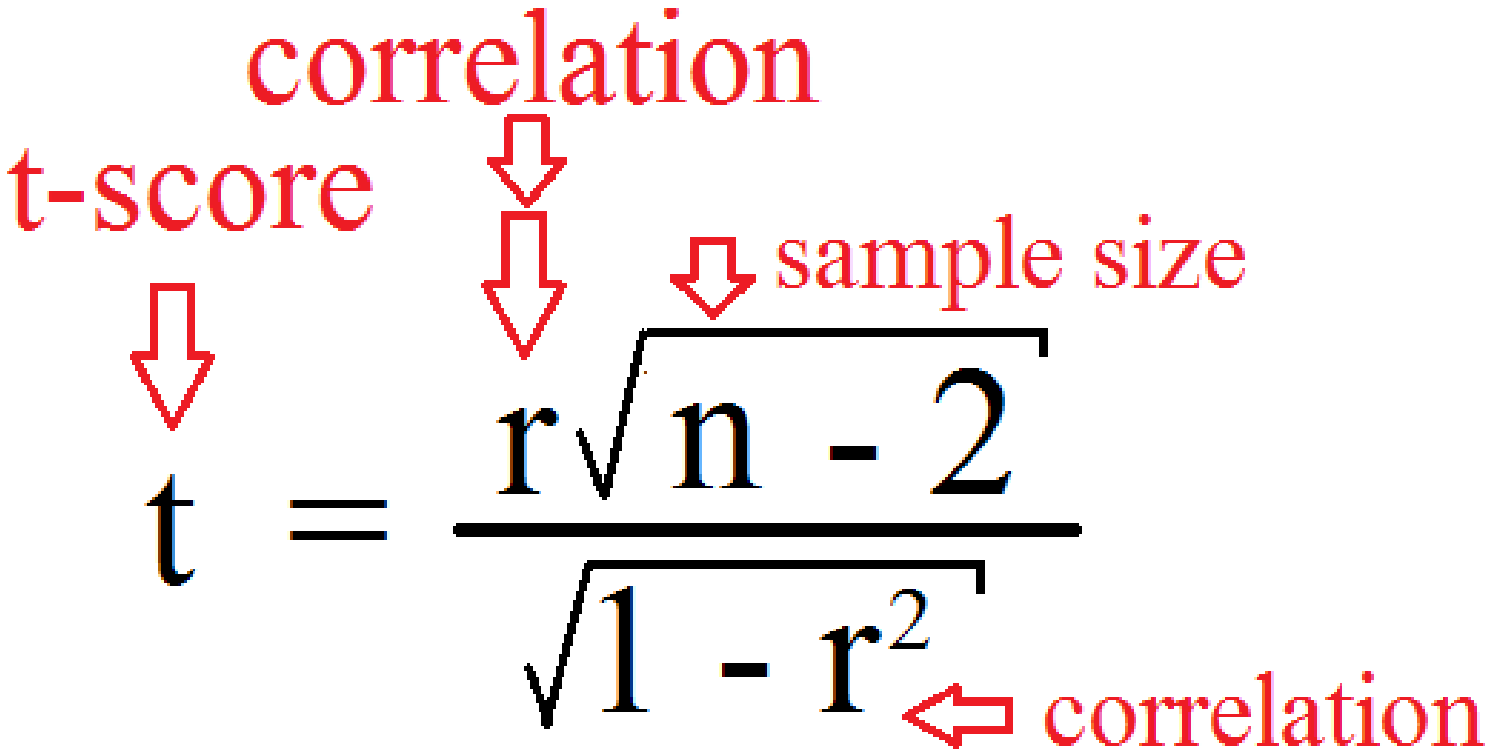
As r gets farther from zero, t-score gets bigger.



So a stronger correlation gives you higher t-score.

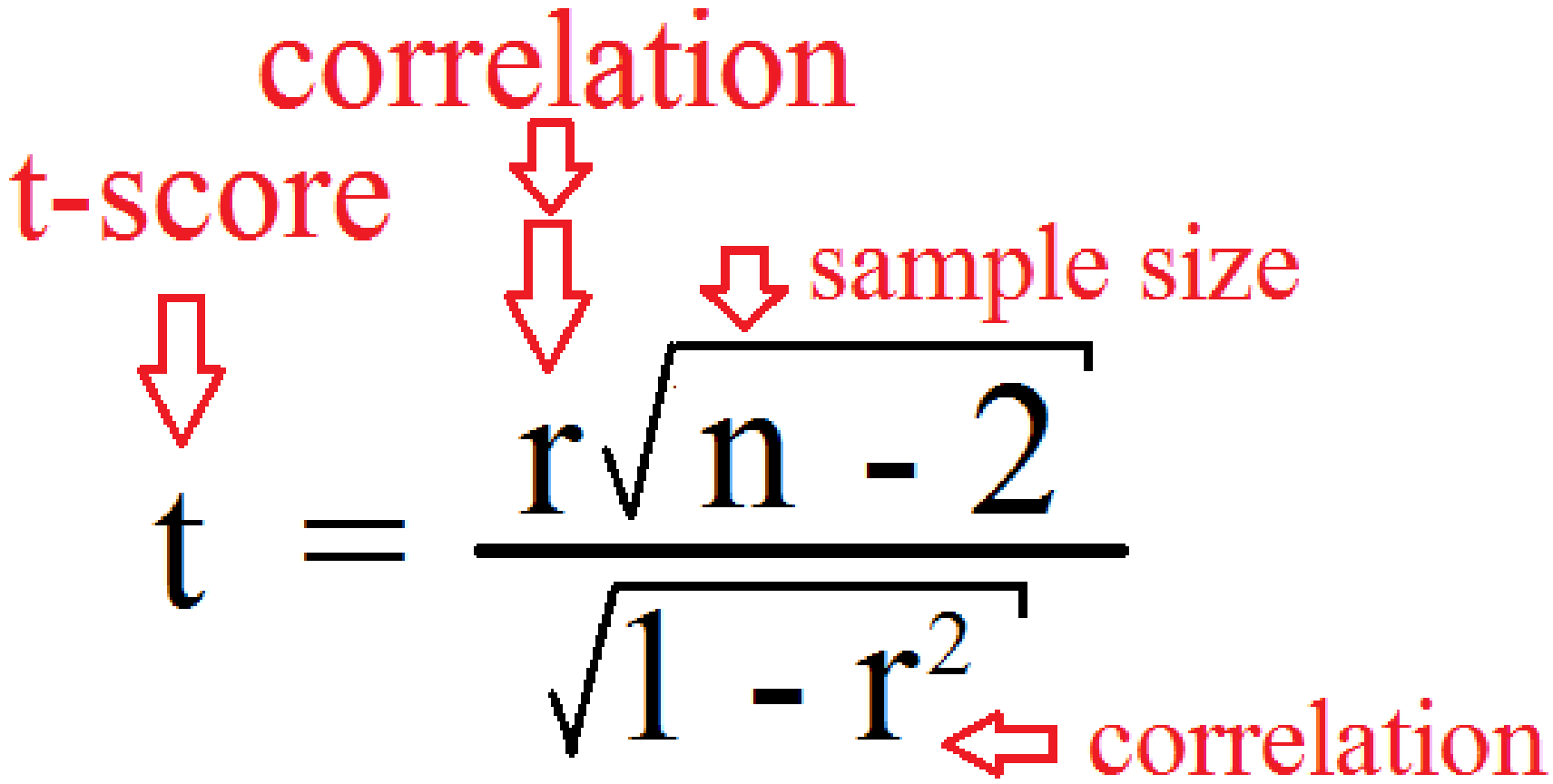
Stronger correlation → better evidence of a correlation.

t-score also increases with sample size. As usual, it's under a square root.

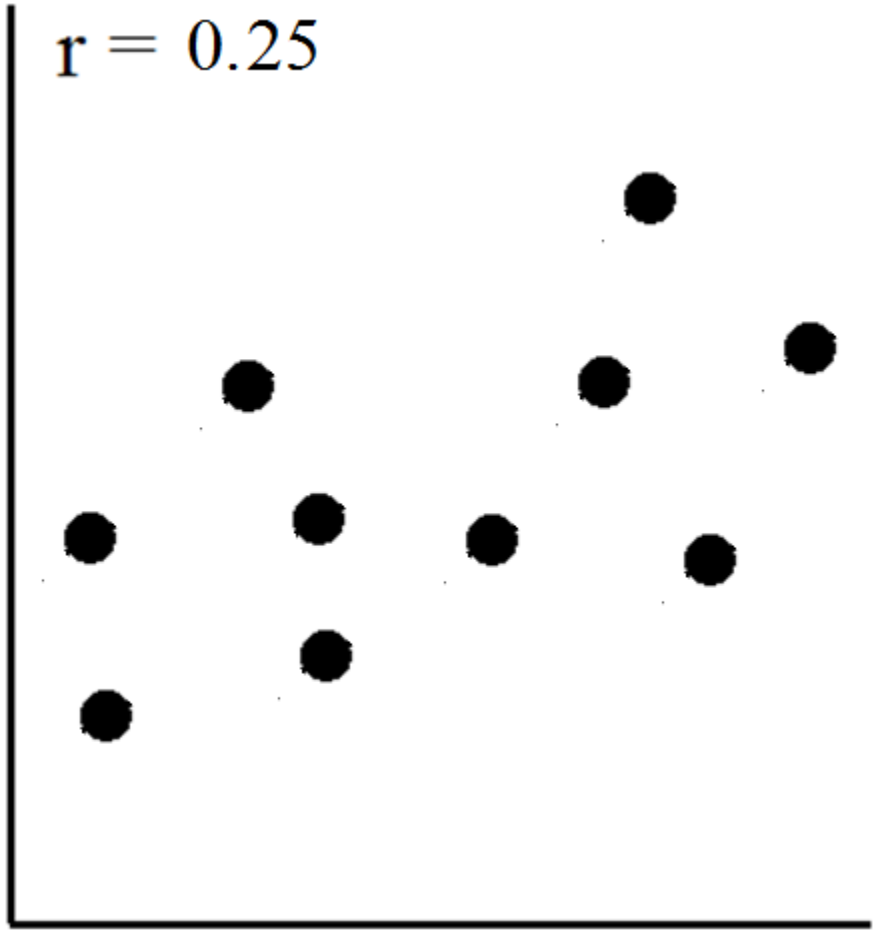


Having more data points makes it easier to detect correlations.

A larger t-score meant more evidence against the null, just like before. So a large t-score means more evidence of a correlation, and of a slope.



If there's a weak correlation and a small sample, we might not detect it. (Example: n=10, r=.25)



$$t = \frac{0.25\sqrt{10 - 2}}{\sqrt{1 - 0.25^2}} = 0.71$$

t* = 1.397, at 8 df, 0.20 significance.

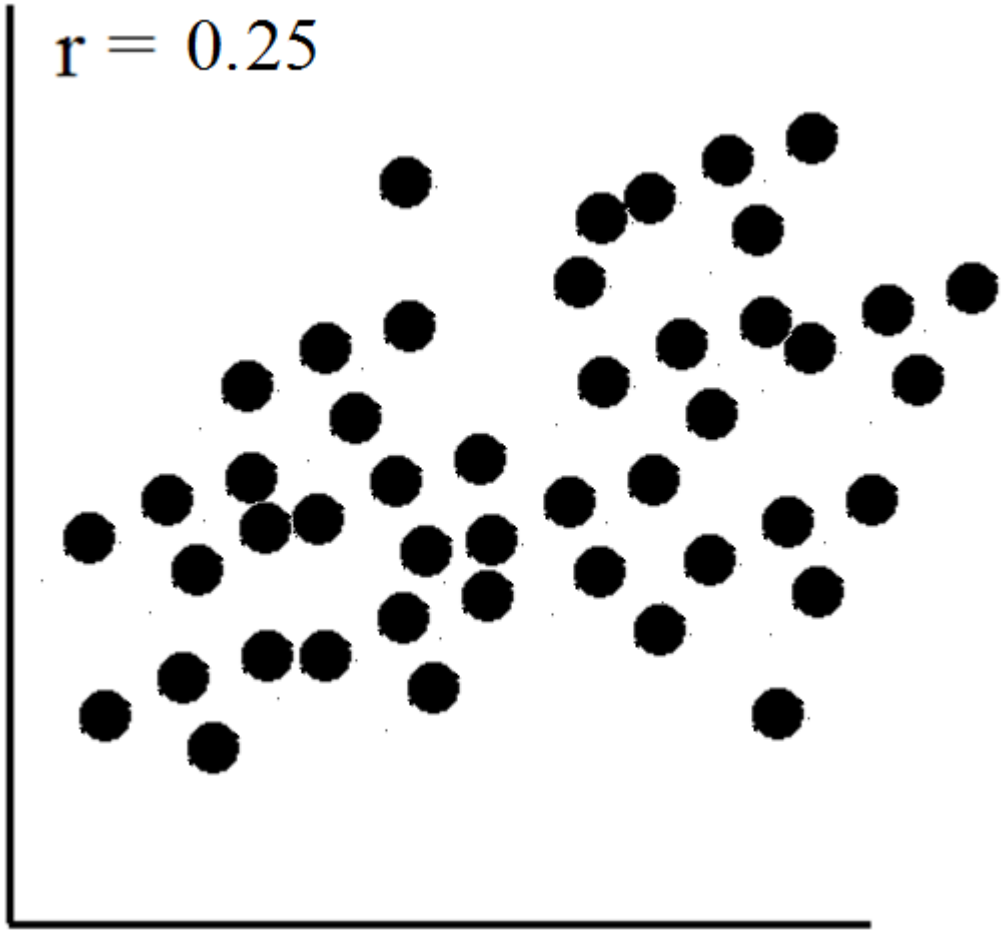
t* = 2.306, at 8 df, 0.05 significance.

The p-value must be ***more than 0.20.***

No evidence of a correlation.

What if we get a larger sample of this correlation?

(n=46, r=0.25)



We should get some evidence of a correlation, but not much.

$$t = \frac{0.25\sqrt{46 - 2}}{\sqrt{1 - 0.25^2}} = 1.71$$

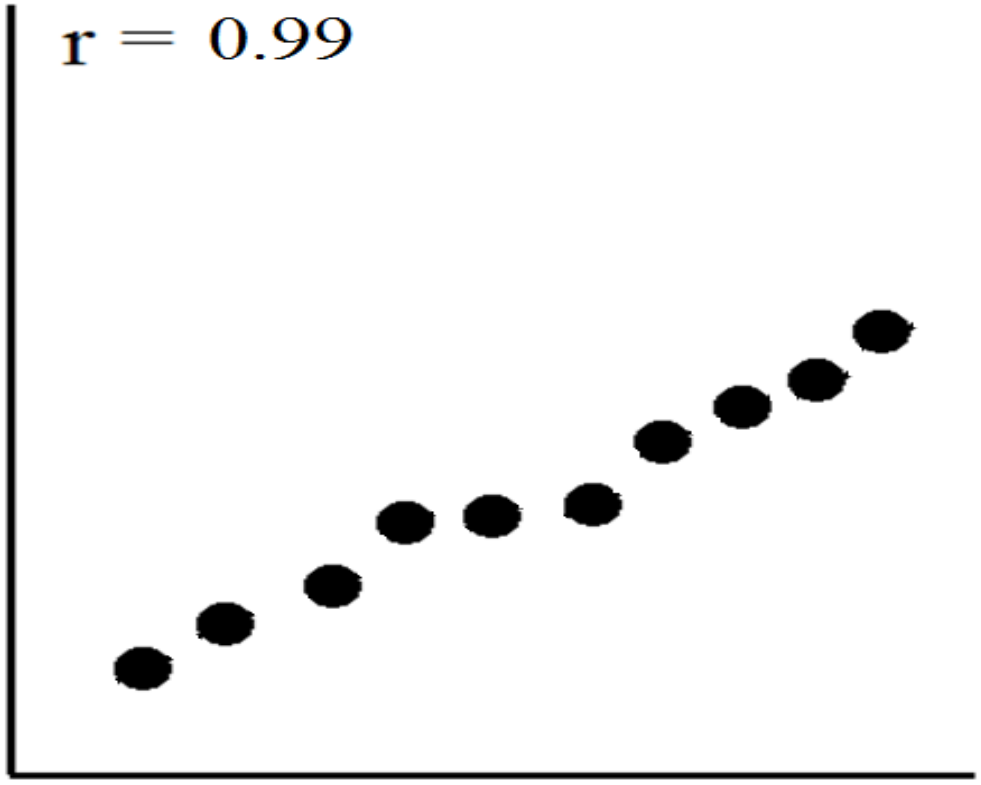
t* = 1.684, at 44 df, 0.10 significance.

t* = 2.021, at 44 df, 0.05 significance.

The p-value must be, ***between 0.05 and 0.10.***

What happens when you get a near perfect correlation?

(Example: n=10, r=.99).



Expectation: Very strong evidence of a correlation.

$$t = \frac{0.99\sqrt{10 - 2}}{\sqrt{1 - 0.99^2}} = 19.85$$

t* = 2.306, at 8 df, 0.05 significance.

t* = 5.041, at 8 df, 0.001 significance.

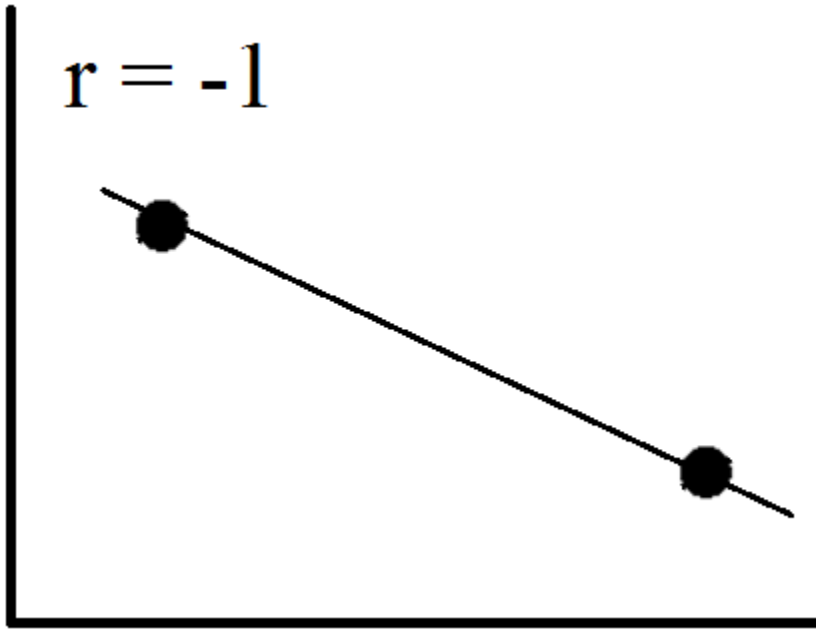
Reality: Very strong evidence of a correlation.

$$t = \frac{0.99\sqrt{10 - 2}}{\sqrt{1 - 0.99^2}} = 19.85$$

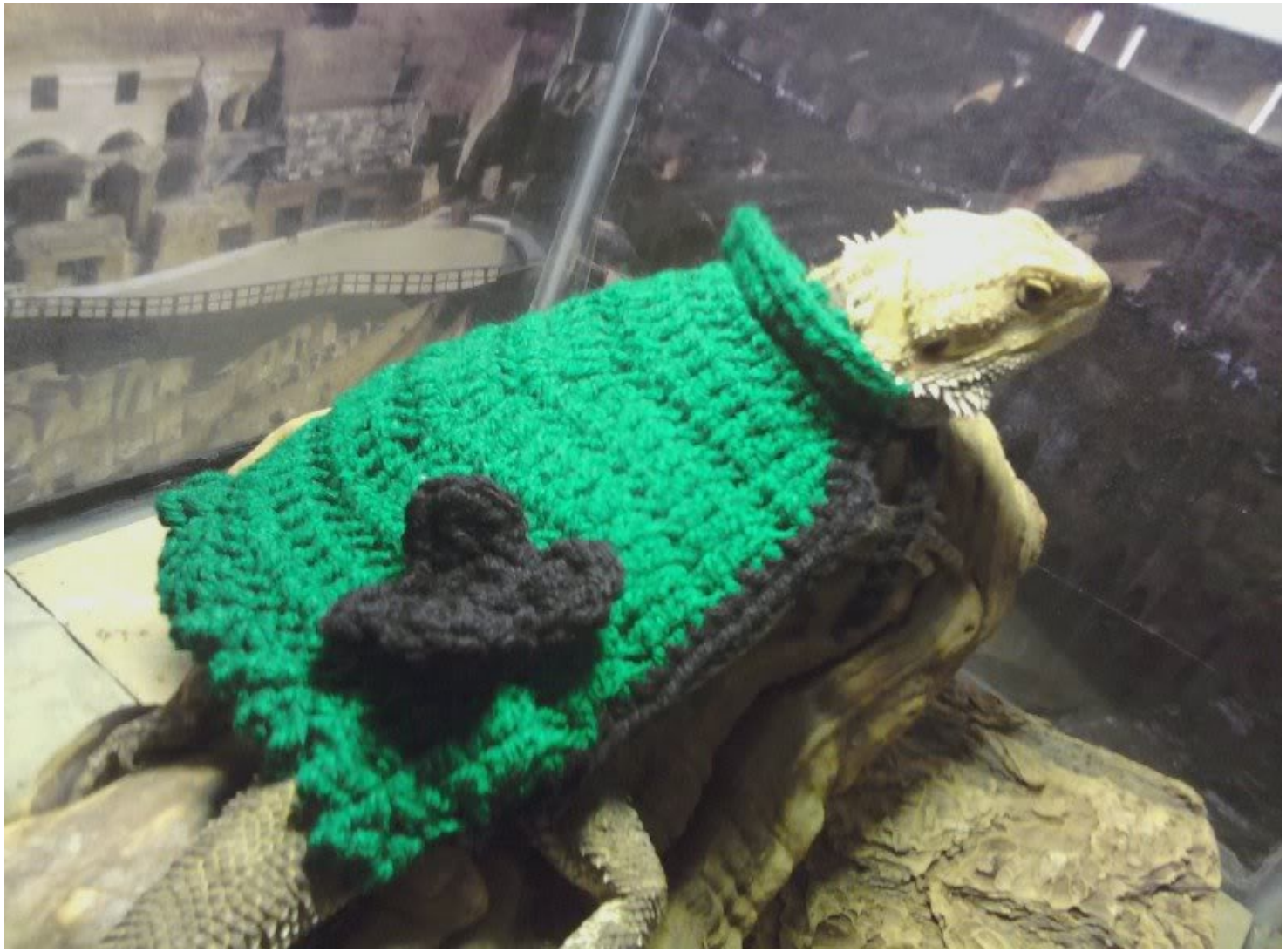
The bottom gets very small, and dividing by a small number gives you something huge.

The same thing happens with a near-perfect negative correlation, but the t-score is negative and huge.

For interest: You can always put a line exactly through two points.



With only two points, we have no idea what the true correlation is. Points after the first two tell us about correlation.



Show your pet some love by forcing it into a tea cozy.

Thursday: **R output, r-squared and the bivariate normal.**

Optional reading note 2.3: Variable names

Some people refer to the x variable as the 'independent' variable, and y as the 'dependent' variable. Don't be one of these people. 'Independent' is already a loaded enough word in stats, and when we get to co-linearity, referring to x as 'independent' will cause headaches.

Diagram used: Ron Howard paper installation from the movie “A Beautiful Mind”
'Monkey Selfie' is part of an ongoing legal battle, fascinating if you're into Copyright Law.
Others recycled from Stat 203 material.