Correlation

Topic 11 · Tuesday, March 2

EDUC 15: Statistics for Education Research
UCI School of Education

Where are we?

Description

Summarize data

Visualize data

Prediction

Does one score go with another?

We are here



Explanation

Peter the Anteater at a UC Irvine basketball game.

Plan for today

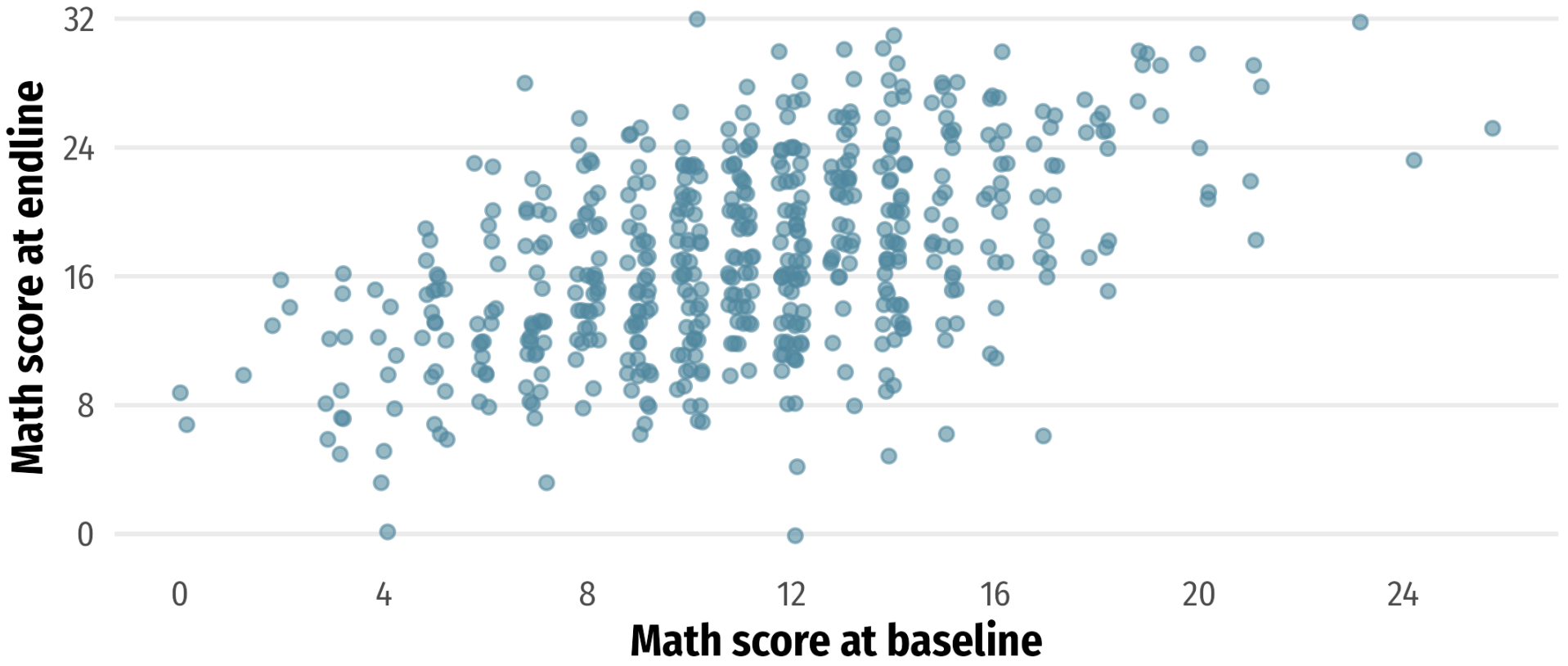
What r shows

Computing r

Testing r

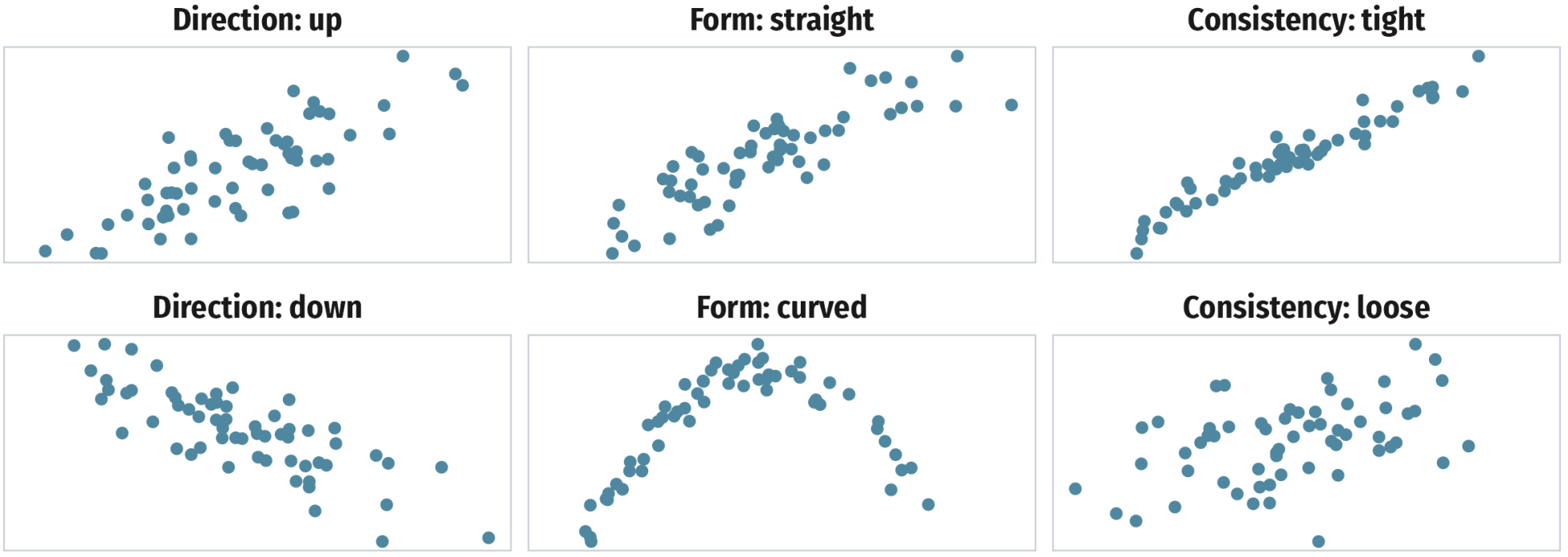
What r shows

Each dot is one student



Data: Muralidharan, Singh, and Ganimian (2019). 539 students. Dots are nudged so equal scores don't hide each other.

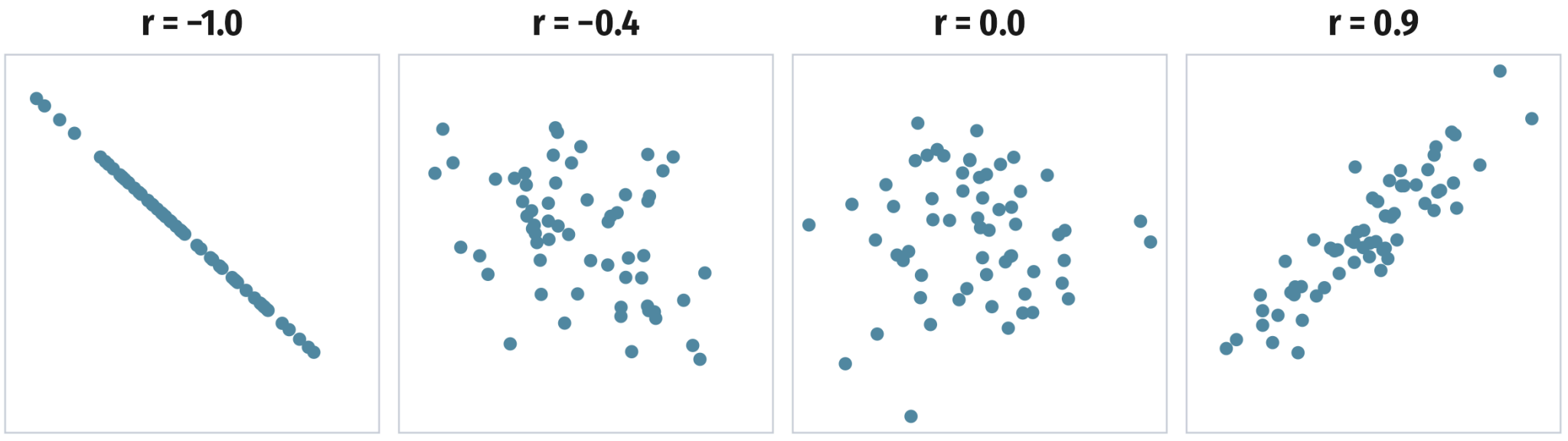
A correlation describes three things



Which way does it go? Is it straight? How closely do the dots follow it?

Made-up data.

r runs from −1 to 1

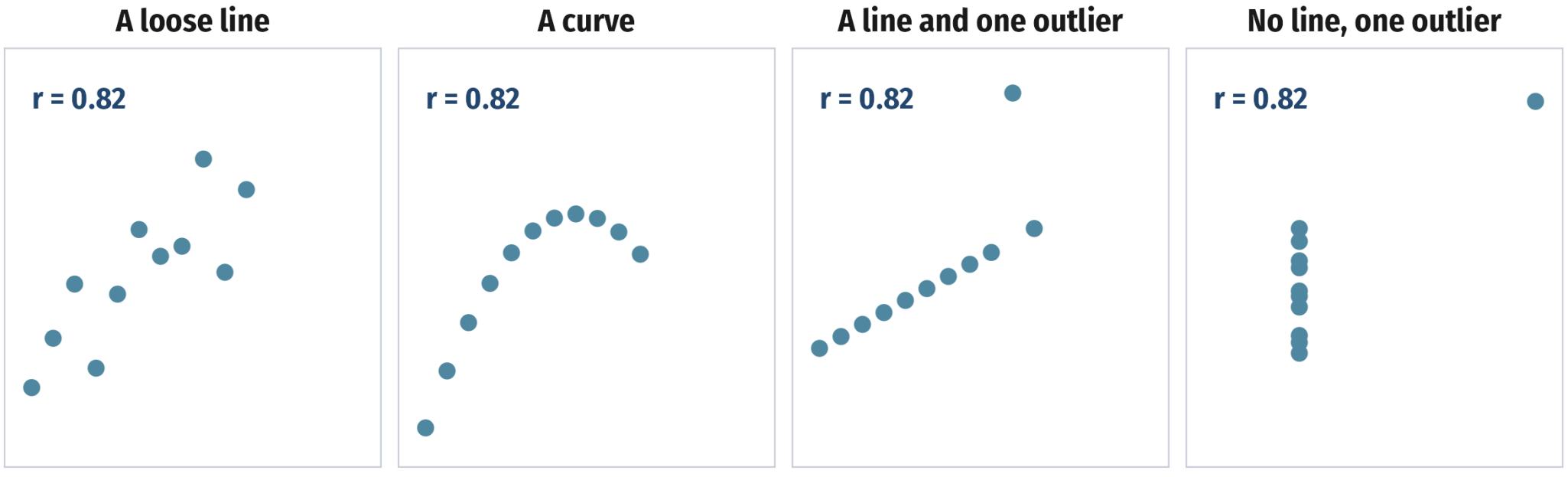


The sign gives the direction. The distance from 0 gives the consistency

Our 539 students: r = 0.50. Upward, roughly straight, and loose

Top: made-up data. Bottom: Muralidharan, Singh, and Ganimian (2019).

Same r, four different stories



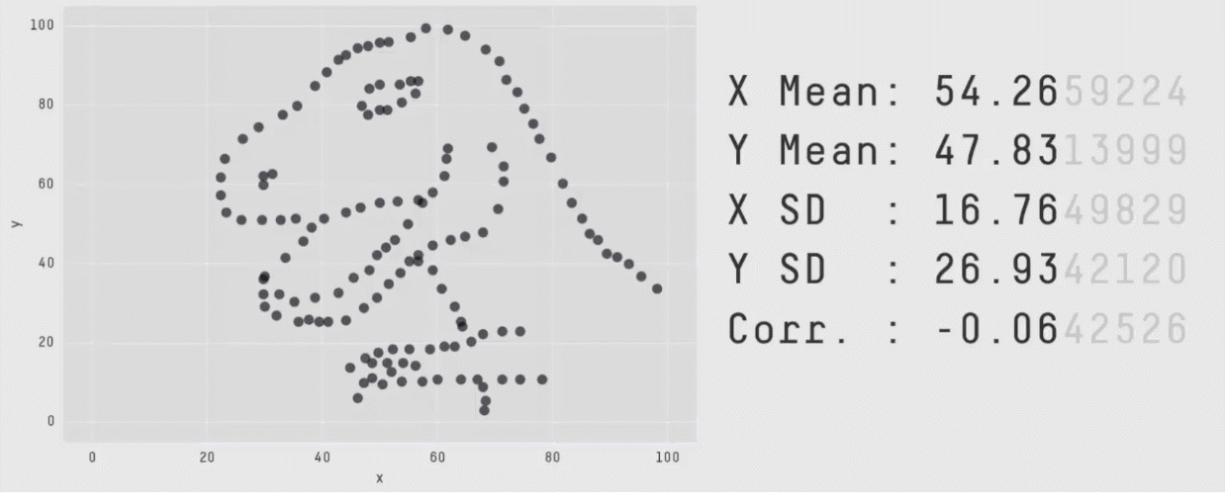
r measures straight lines only. The second panel is a perfect curve, and r can't tell

Plot your data before you trust the number

Data: Anscombe (1973), four made-up data sets.

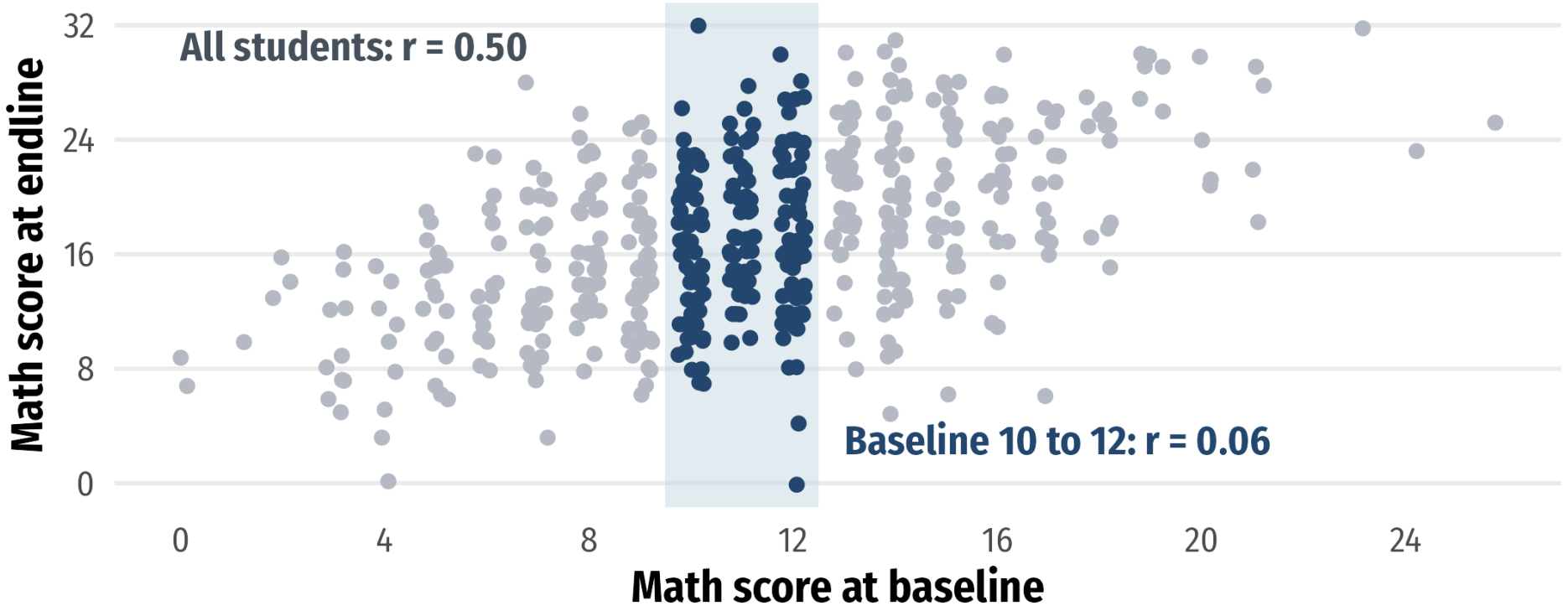
Don't forget to plot your data

All of these data sets have the same means, SDs and correlation



Matejka and Fitzmaurice (2017), "Same Stats, Different Graphs," CHI 2017. Datasaurus: Alberto Cairo.

A narrow range hides a correlation



Same tests, 181 students. We can only speak for the range we observed

Data: Muralidharan, Singh, and Ganimian (2019).

Predict it: One point, far from the rest

Respond at **PollEv.com/adb**

A scatter plot has 12 points and r = 0.60. We add one new point: far to the right, and near the bottom. What happens to r?

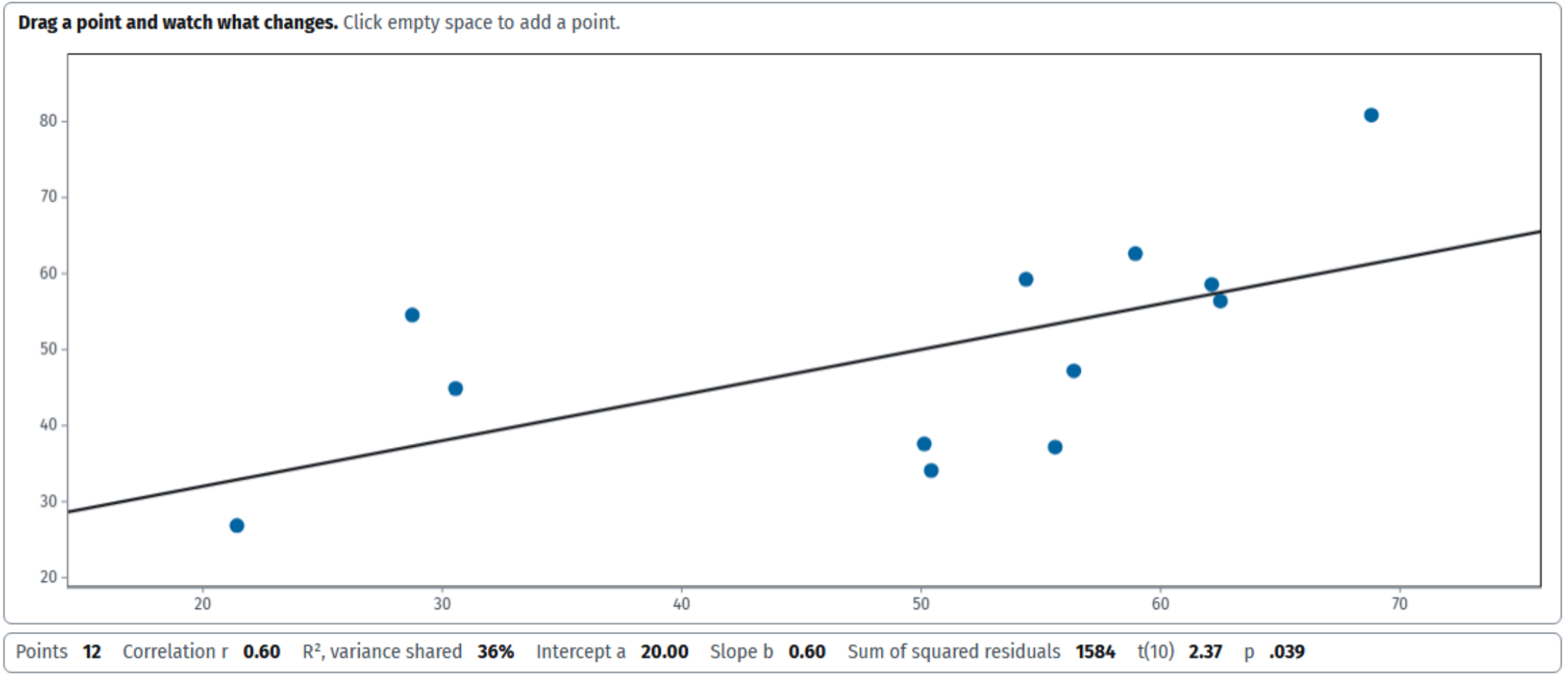
A It goes up

B It barely moves

C It drops a lot

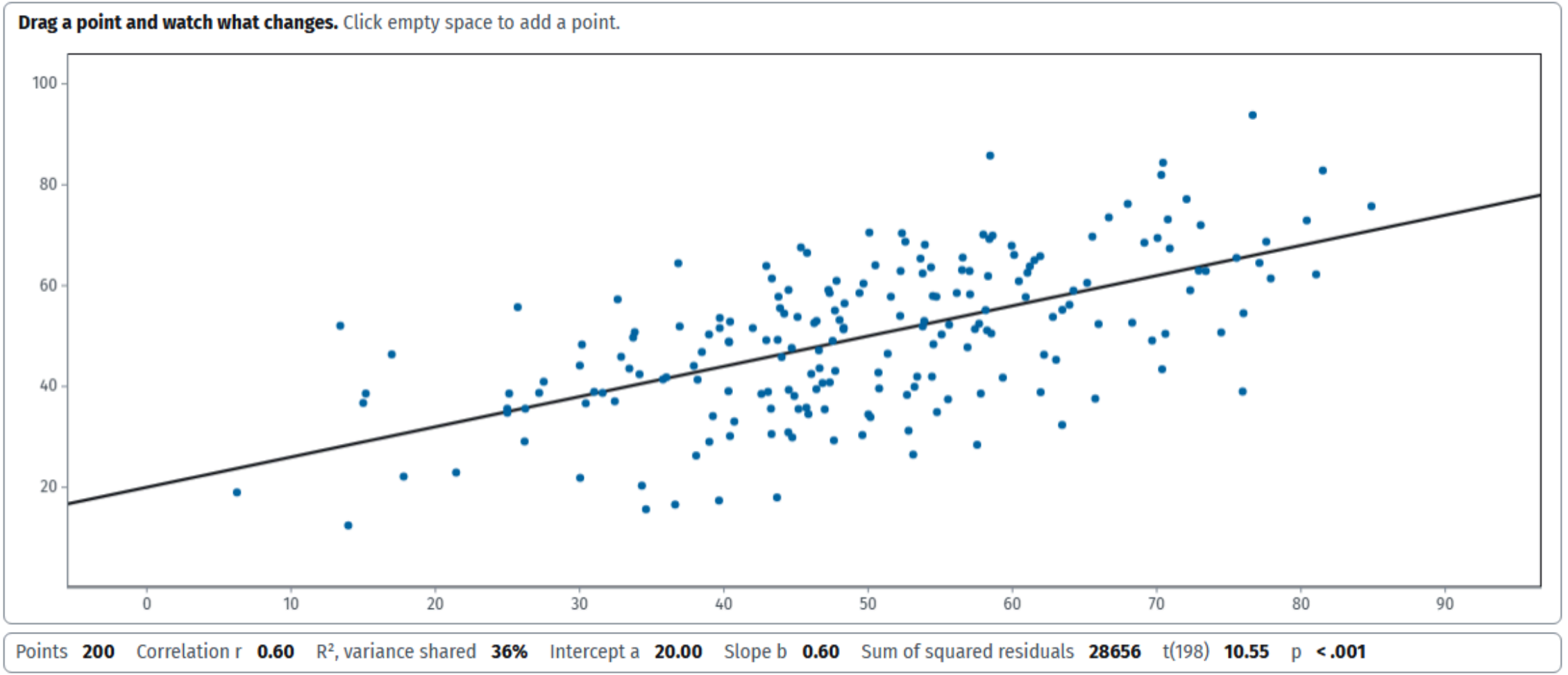
D It becomes exactly 0

Predict it: Add the outlier



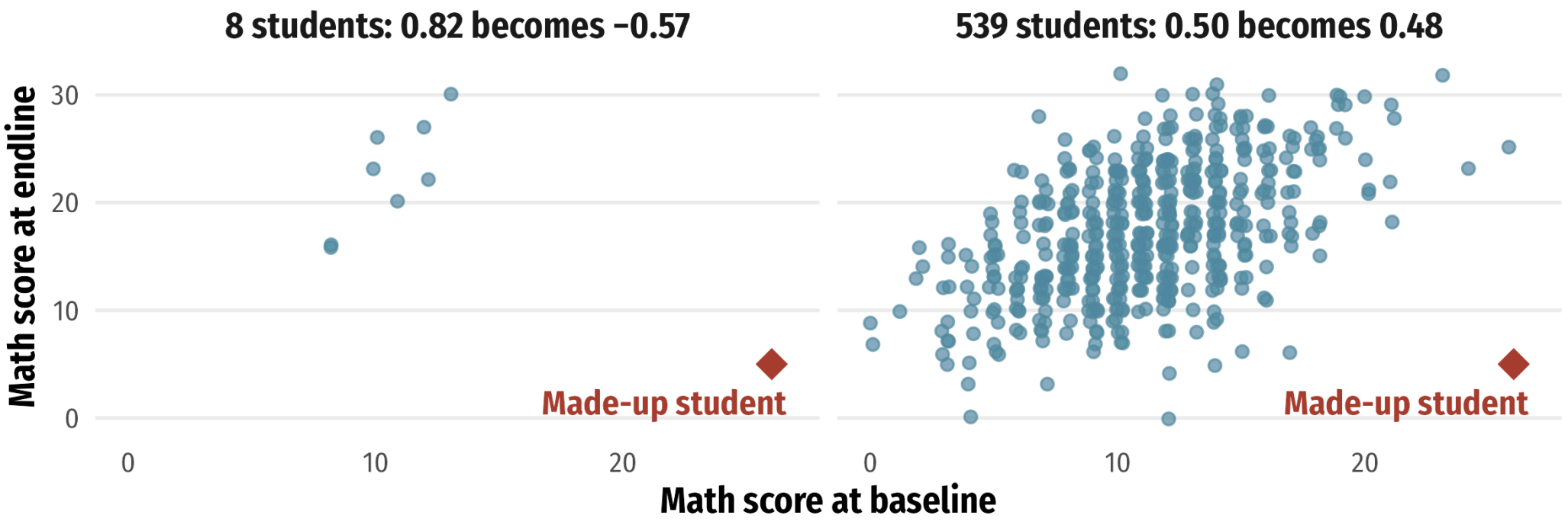
Simulator after Kristoffer Magnusson, rpsychologist.com. On your phone: the simulator page.

Predict it: Same move, 200 points



Simulator after Kristoffer Magnusson, rpsychologist.com.

One outlier swings 8 students, not 539



The same made-up student flips the sign for 8 students, and barely touches the 539. Outliers matter less as n grows

One made-up student added. Other data: Muralidharan, Singh, and Ganimian (2019).

So what is a correlation good for?

Prediction

Baseline scores can flag who may need help, whatever the reason

Validity

A new math test should correlate with an established one

Reliability

The same test, given twice, should give similar scores

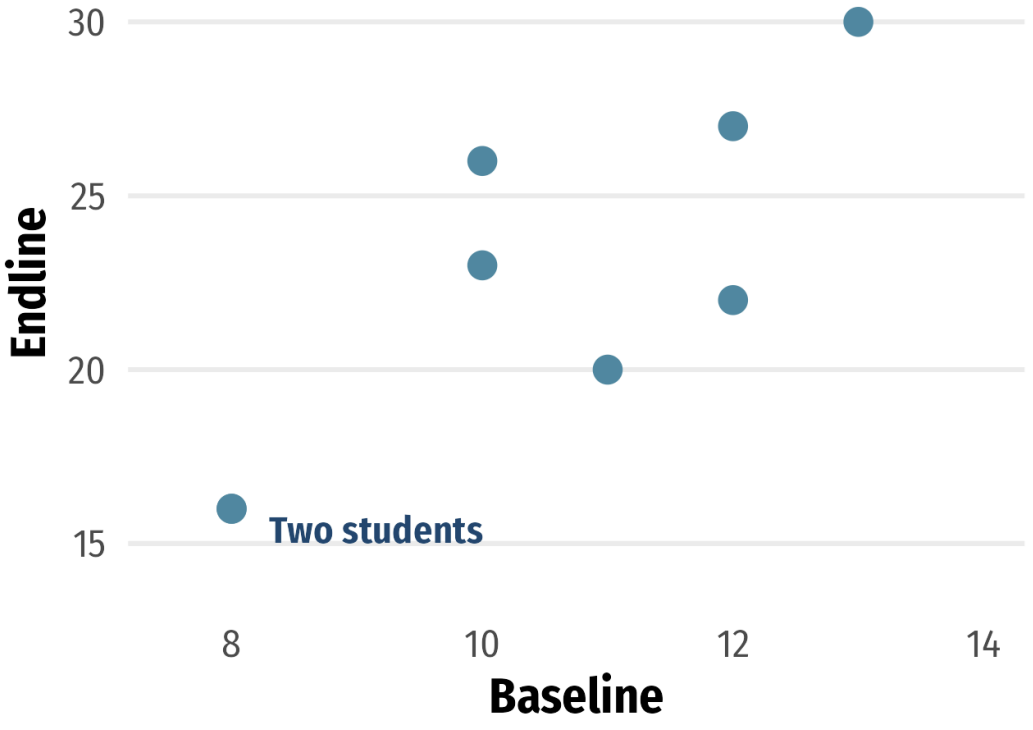
Testing theories

A theory says two things go together. Do they?

Computing r

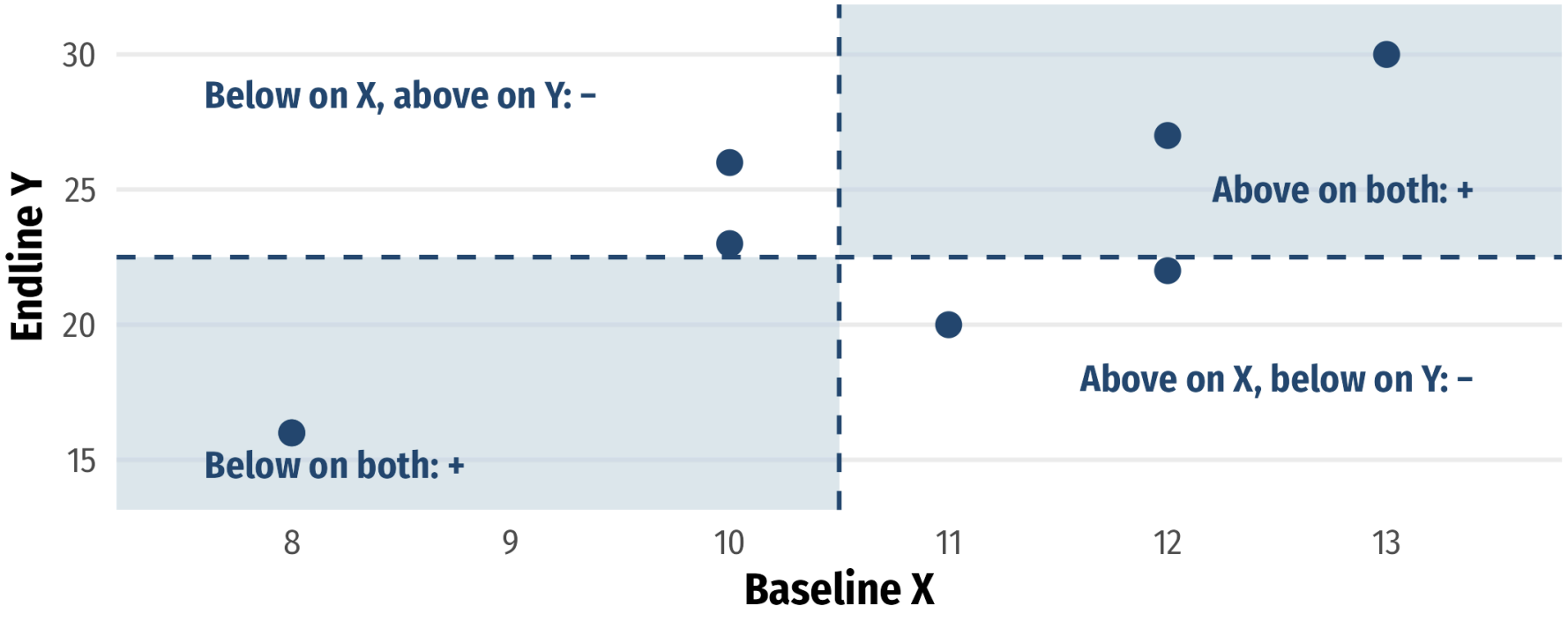
Eight students, by hand

Baseline X	Endline Y
13	30
11	20
8	16
10	26
12	27
8	16
12	22
10	23



Data: Muralidharan, Singh, and Ganimian (2019). Students 25 to 32.

Do the two scores move together?



Multiply each student's two deviations. Same side of both means gives a positive product

Data: Muralidharan, Singh, and Ganimian (2019).

The sum of products, SP

X	Y	$X - M_X$	$Y - M_Y$	Product
13	30	2.5	7.5	18.75
11	20	0.5	−2.5	−1.25
8	16	−2.5	−6.5	16.25
10	26	−0.5	3.5	−1.75
12	27	1.5	4.5	6.75
8	16	−2.5	−6.5	16.25
12	22	1.5	−0.5	−0.75
10	23	−0.5	0.5	−0.25
			SP	**54.00**

Shortcut: $SP = \Sigma XY - (\Sigma X)(\Sigma Y) / n = 1{,}944 - (84 \times 180) / 8 = 54$

Data: Muralidharan, Singh, and Ganimian (2019).

r divides by each score's own spread

$$r = \frac{SP}{\sqrt{SS_X \, SS_Y}}$$

Top: how much X and Y move together

Bottom: how much each moves alone

$$r = \frac{54}{\sqrt{24 \times 180}} = \frac{54}{65.73} = 0.822$$

The bottom is why r has no units, and why it stays between −1 and 1

r^2 is the share we can predict

Group	r	r^2	In words
Our 8 students	0.822	0.675	67.5% of Y's variability can be predicted from X
All 539 students	0.502	0.252	25.2% of Y's variability can be predicted from X

r = 0.50 doesn't mean half. It means about a quarter of the variability is shared

r^2 works like $\hat{\eta}^2$ in ANOVA: a share of the total. It's our effect size

In the book: r^2 is the coefficient of determination

Data: Muralidharan, Singh, and Ganimian (2019).

Try it: Five more students

Baseline X	Endline Y
13	14
10	21
11	24
11	14
10	22

$M_X = 11, M_Y = 19$

$SS_X = 6, SS_Y = 88$

Find SP, then r

Work alone, compare with a neighbor, then vote with hands: is r positive or negative?

Data: Muralidharan, Singh, and Ganimian (2019).

Try it: Answer

$X - M_X$	$Y - M_Y$	Product
2	−5	−10
−1	2	−2
0	5	0
0	−5	0
−1	3	−3

SP = −15

$r = -15 / \sqrt{(6 \times 88)} = -0.65$

Eight students said 0.82, and these five say −0.65. All 539 say 0.50. Who's right?

Data: Muralidharan, Singh, and Ganimian (2019).

Testing r

Three samples, three values of r

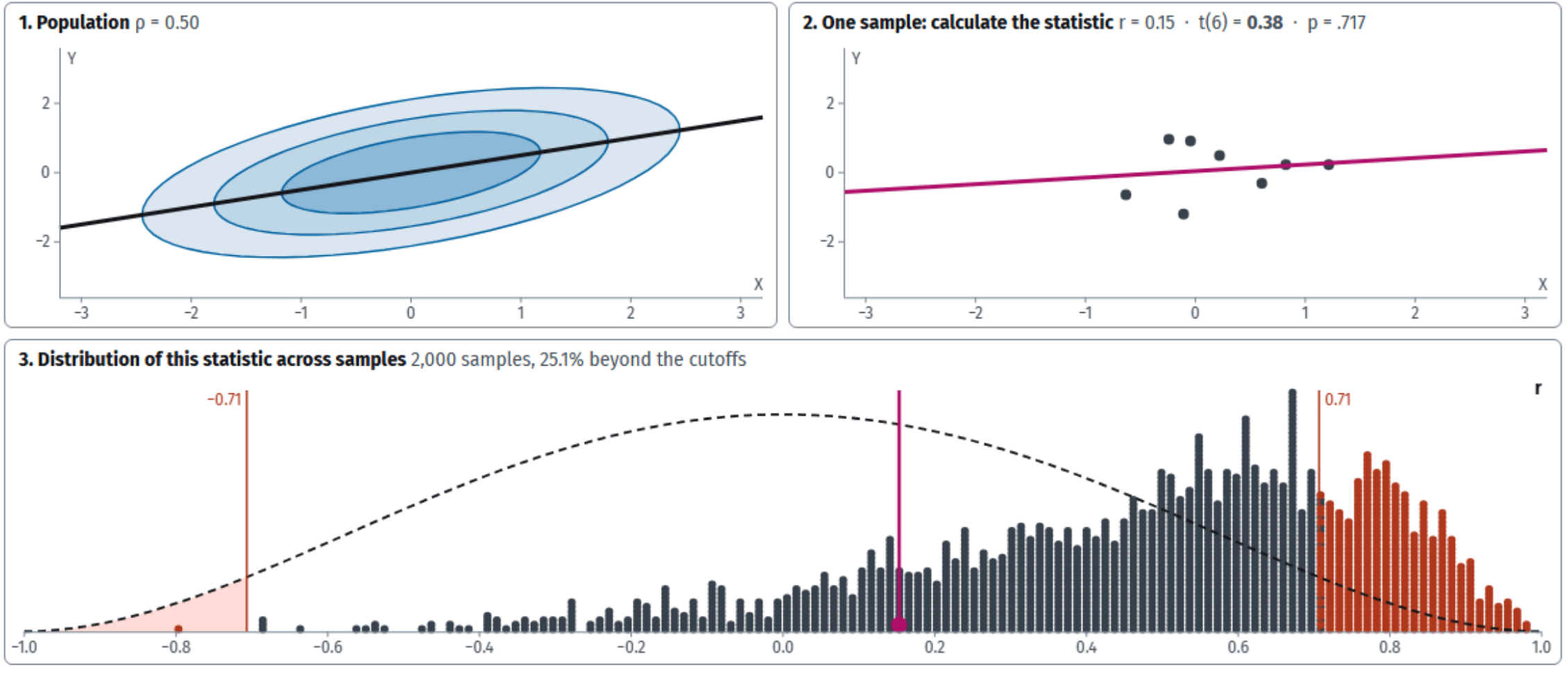
Students	n	r
Students 25 to 32	8	0.82
The five from the try-it	5	−0.65
Everyone with both scores	539	0.50

Each r estimates one number: ρ, the correlation in the population

Greek means population. r is to ρ what M is to μ

Data: Muralidharan, Singh, and Ganimian (2019).

With 8 students, r bounces a lot



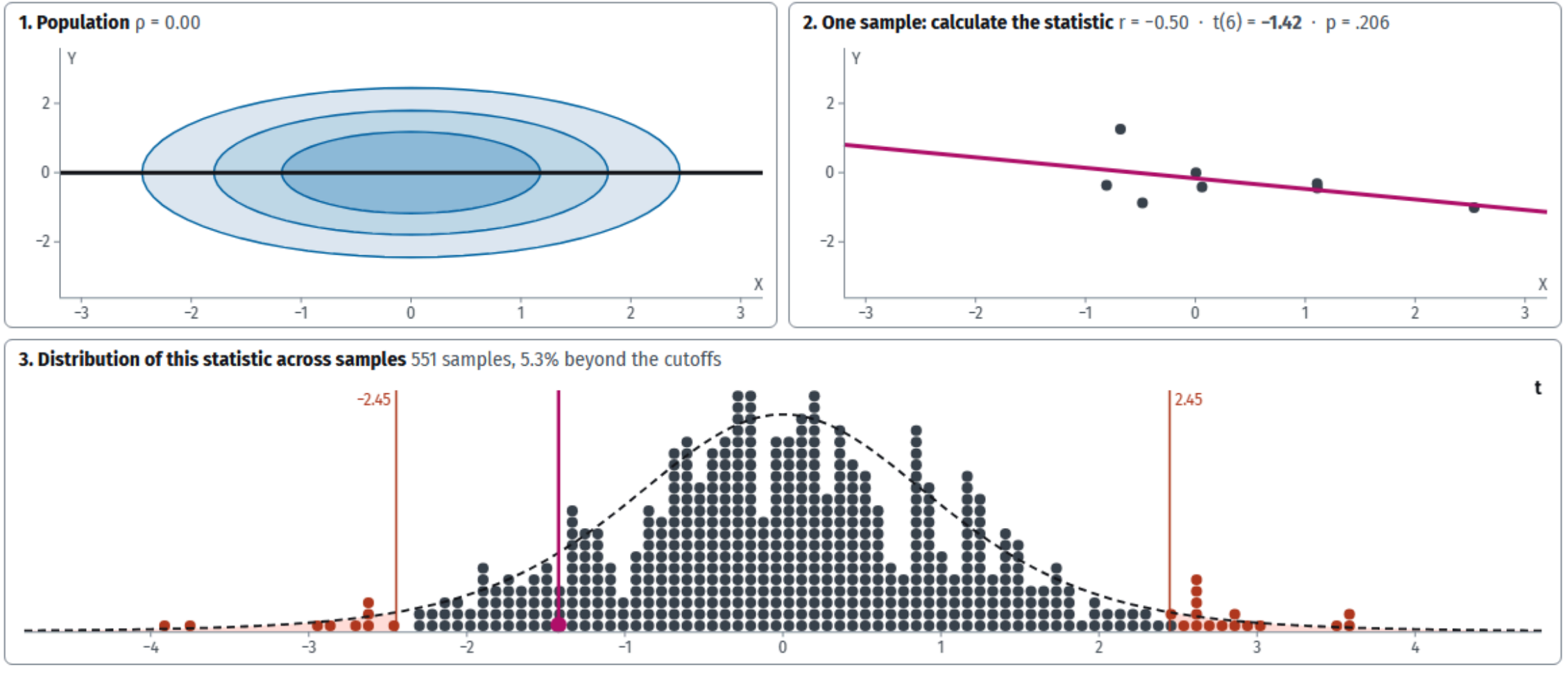
Simulator after Kristoffer Magnusson, rpsychologist.com.

Steps 1 and 2: H_0 says $\rho = 0$

Step	In words	Our 8 students
1. State H_0	No relationship in the population	H_0: $\rho = 0$ H_1: $\rho \neq 0$
2. Set the criteria	Two tails, $\alpha = 0.05$, and a t curve	$df = n - 2 = 6$ t(6) beyond ±2.447

The rule stays the same: is what we saw surprising, if nothing is going on?

Under H_0, r still bounces



Simulator after Kristoffer Magnusson, rpsychologist.com.

Step 3: The same fraction again

$$t = \frac{r - \rho}{\hat{\sigma}_r} \qquad \text{where} \qquad \hat{\sigma}_r = \sqrt{\frac{1 - r^2}{n - 2}}$$

Top: how far r is from H_0's ρ, which is 0

Bottom: how far an r typically falls

Ours: $\hat{\sigma}_r = \sqrt{(0.325 / 6)} = 0.233$, so $t = 0.822 / 0.233 = 3.53$

In the book: $\hat{\sigma}_r$ is written s_r

Eight students or 539: reject H_0

Step	Our 8 students	All 539 students
1. State H_0	$\rho = 0$	$\rho = 0$
2. Set the criteria	$t(6)$ beyond ± 2.447	$t(537)$ beyond ± 1.964
3. Find the spread	$\hat{\sigma}_r = 0.233$	$\hat{\sigma}_r = 0.037$
3. Compute t	$t = 3.53$	$t = 13.46$
4. Decide	Reject H_0: $3.53 > 2.447$	Reject H_0: $13.46 > 1.964$

Eight students: $t(6) = 3.53$, $p = 0.012$, $r = 0.82$

All 539: $t(537) = 13.46$, $p < 0.001$, $r = 0.50$, 95% CI [0.44, 0.56]

Data: Muralidharan, Singh, and Ganimian (2019). Software gives the interval for r.

A correlation can't name a cause

Pair	r	t	Who set X?	A cause?
Tuition, baseline math	0.10	t(617) = 2.43	Families	We can't say
Lottery, endline math	0.23	t(537) = 5.37	A lottery	Yes

Both correlations are significant. Math scores might drive tuition, or a third thing might drive both

The lottery is the difference. It broke every other link, so only the program is left

Data: Muralidharan, Singh, and Ganimian (2019).

Other kinds of scores, same formula

The two variables	Name	What it does	Mindspark
Both continuous	Pearson's r	What we did today	Baseline, endline: 0.50
One or both ordinal	Spearman's r_S	Pearson's r on ranks	Same scores, ranked: 0.48
One binary, one continuous	Point-biserial r	Pearson's r, with 0 and 1	Lottery, endline: 0.23
Both binary	Phi, φ	Pearson's r, with 0 and 1 for both	Lottery, top half at endline: 0.19

The lottery's t(537) = 5.37 is Topic 9's two-sample t-test. Same data, same t

Data: Muralidharan, Singh, and Ganimian (2019).

Try it: Nine fifth graders

Nine students started in grade 5. Their baseline and endline scores have r = 0.583

Test H_0: $\rho = 0$ with $\alpha = 0.05$ and two tails. The cutoff is $t(7) = \pm 2.365$

Find $\hat{\sigma}_r$ and t, then decide. Work alone, compare with a neighbor, then vote with hands

This r is bigger than the 0.50 from all 539. Does that make it easier to reject?

Data: Muralidharan, Singh, and Ganimian (2019).

Try it: Answer

$\hat{\sigma}_r = \sqrt{((1 - 0.340) / 7)} = \mathbf{0.307}$

$t = 0.583 / 0.307 = \mathbf{1.90}$

1.90 is inside ±2.365, so we maintain H_0: $t(7) = 1.90$, $p = 0.099$

A bigger r than everyone's, and still no rejection. With n = 9, an r this size isn't surprising if ρ = 0

Maintaining H_0 doesn't mean ρ = 0. It means 9 students can't tell

Data: Muralidharan, Singh, and Ganimian (2019).

The blueprint so far

Blueprint step	Statistic	Bottom of the fraction	df
1. One score	z	σ	None
2. One mean, σ known	z	σ_M	None
3a. One mean, σ unknown	t	$\hat{\sigma}_M$	$n - 1$
3b. Mean gain	t	$\hat{\sigma}_M$ of the gains	$n - 1$
4. Two means	t	$\hat{\sigma}_{M_1 - M_2}$	$n_1 + n_2 - 2$
5. Two or more means	F	MS_{within}	$k - 1$ and $N - k$
6. A correlation	**t**	$\hat{\sigma}_r$	**$n - 2$**
7a. A slope: next	t	Topic 13	

Same idea every time: what we saw, over what chance alone would give

What you can do now

Read direction, form and consistency from a scatter plot

Say what outliers, curves and a narrow range do to r

Compute SP, r and r^2 by hand

Test H_0: $\rho = 0$ with t and $n - 2$ degrees of freedom

Explain why a correlation alone doesn't show a cause

Next steps

Thursday, March 4: Topic 12

Regression: the line. Read the first part of section 14-6

Friday, March 5: Lab 9

Correlation and regression. Bring a calculator and a laptop

Mastery and Problem Set for Chapter 14 are due Sunday, March 7 at 6 PM