t-tests for two samples and for gains

Topic 9 · Thursday, February 18 and Tuesday, February 23

EDUC 15: Statistics for Education Research
UCI School of Education

Where are we?

Description

Summarize data

Visualize data

Prediction

Explanation

Did the program cause the gap?

We are here



Peter the Anteater at a UC Irvine basketball game.

The four steps stay, for a gap

Step 1: State a hypothesis about the difference between two populations

Step 2: Set the criteria: what's unexpected among differences?

Step 3: Observe a sample from each population, and compute their difference

Step 4: Reject or maintain H_0

Plan for today

From one mean to two

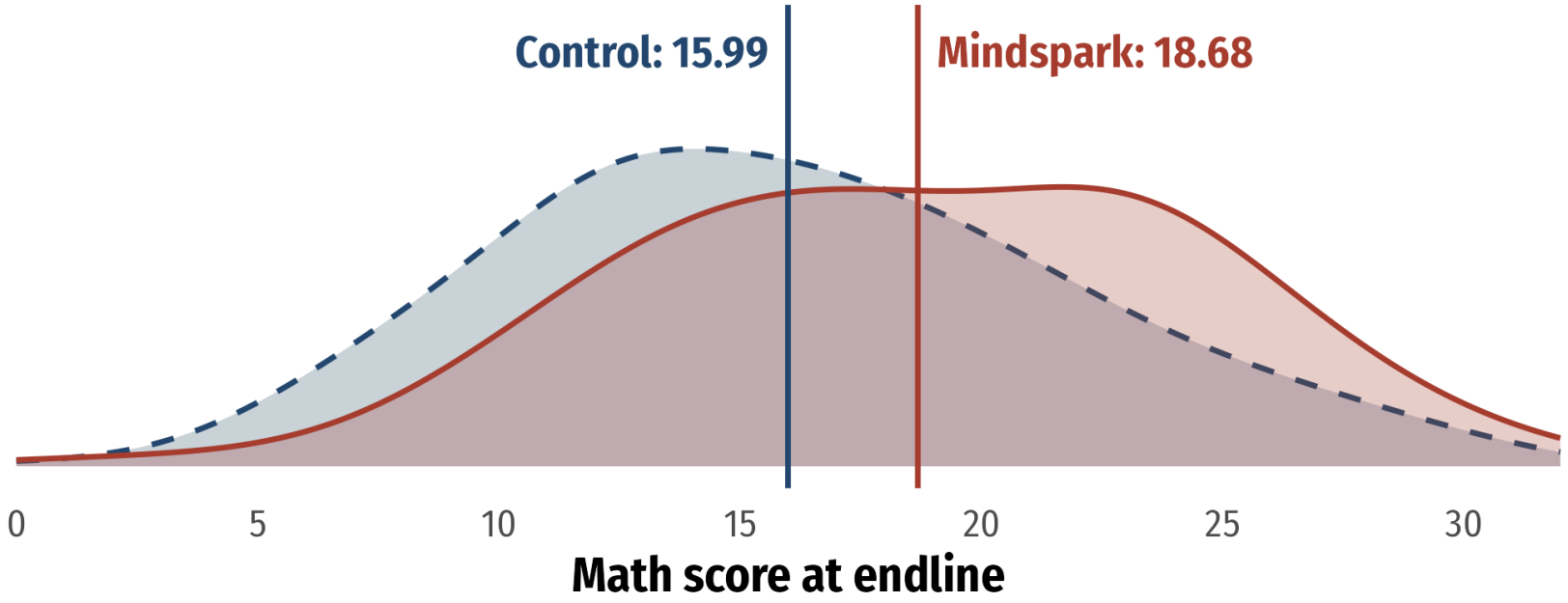
The two-sample t

Mindspark's test

On Tuesday: what if the same people take the test twice?

From one mean to two

Lottery winners scored higher



The gap is 18.68 − 15.99 = 2.69 points. Could that be sampling error?

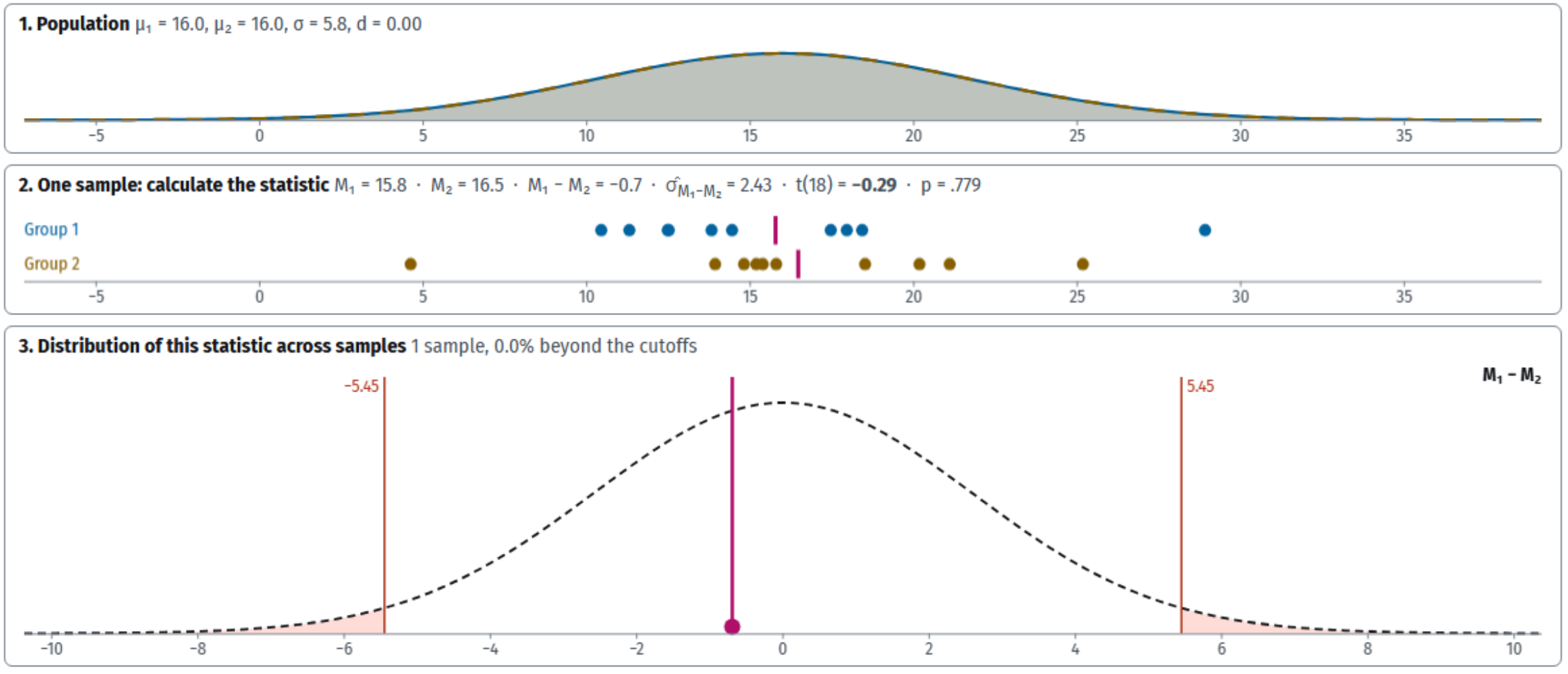
Data: Muralidharan, Singh, and Ganimian (2019). 266 Mindspark students and 273 control students.

The blueprint gets a new step

Blueprint step	Panel 2: one sample gives	Panel 3: the pile of
1. One score	z	z-scores
2. One mean, σ known	z	z-scores of means
3a. One mean, σ unknown	t	t-scores of means
4. Two means	**t for $M_1 - M_2$**	**t-scores of differences**

Same three panels: population, one sample, the statistic across samples

Step 4: two populations, two samples



Simulator after Kristoffer Magnusson, rpsychologist.com.

H_0 now says the gap is zero

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_1 : \mu_1 - \mu_2 \neq 0$$

H_0: the two population means are the same. Across samples, gaps average zero

H_1: the two populations differ. Gaps pile up around some other value

A gap of 2.69 under H_0 can only be sampling error. How unusual would that be?

Predict it: How far does a gap fall?

Respond at **PollEv.com/adb**

H_0 is true, with $\sigma = 5.8$ and $n = 10$ in each group. One group's mean typically falls 1.83 points from its μ. How far does the gap $M_1 - M_2$ typically fall from 0?

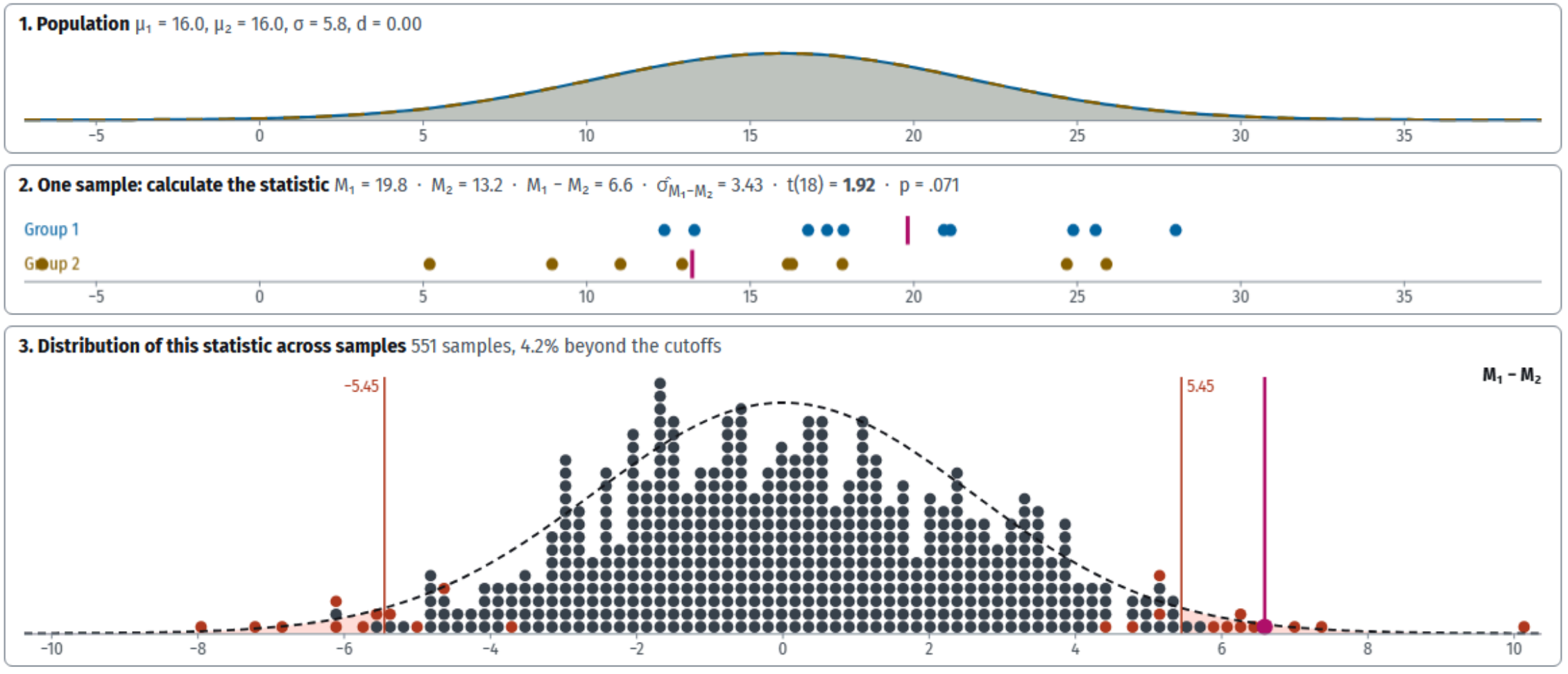
A About 0.9: half as far

B About 1.8: the same

C About 2.6: a bit farther

D About 3.7 or more: at least twice as far

The gaps pile up around zero



Simulator after Kristoffer Magnusson, rpsychologist.com.

A gap has its own standard error

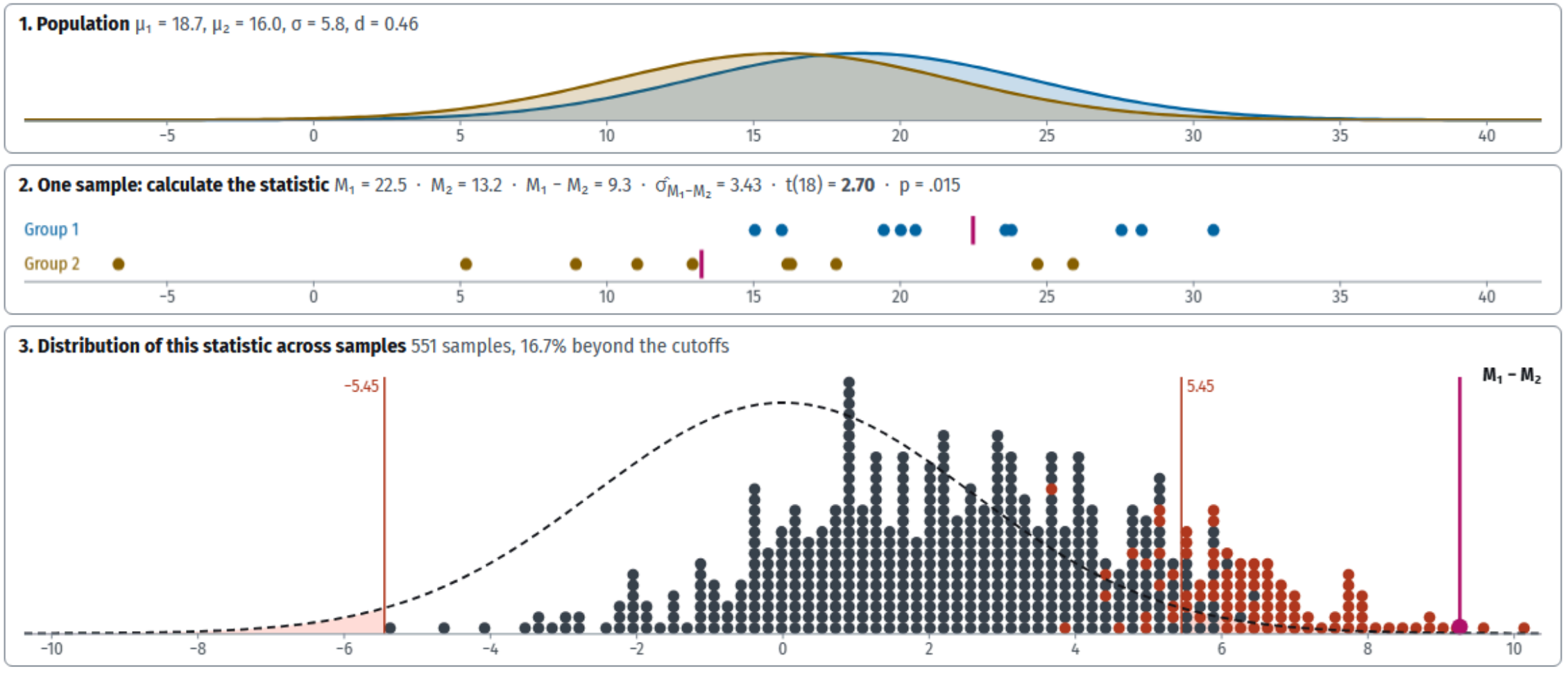
$$\sigma_{M_1 - M_2} = \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}$$

Variances add, even when the means subtract. Each sample brings its own error

Here: $\sqrt{(33.64 / 10 + 33.64 / 10)} = 2.59$. With n = 200 each, it's 0.58

We're assuming both populations share one σ. Keep that in mind

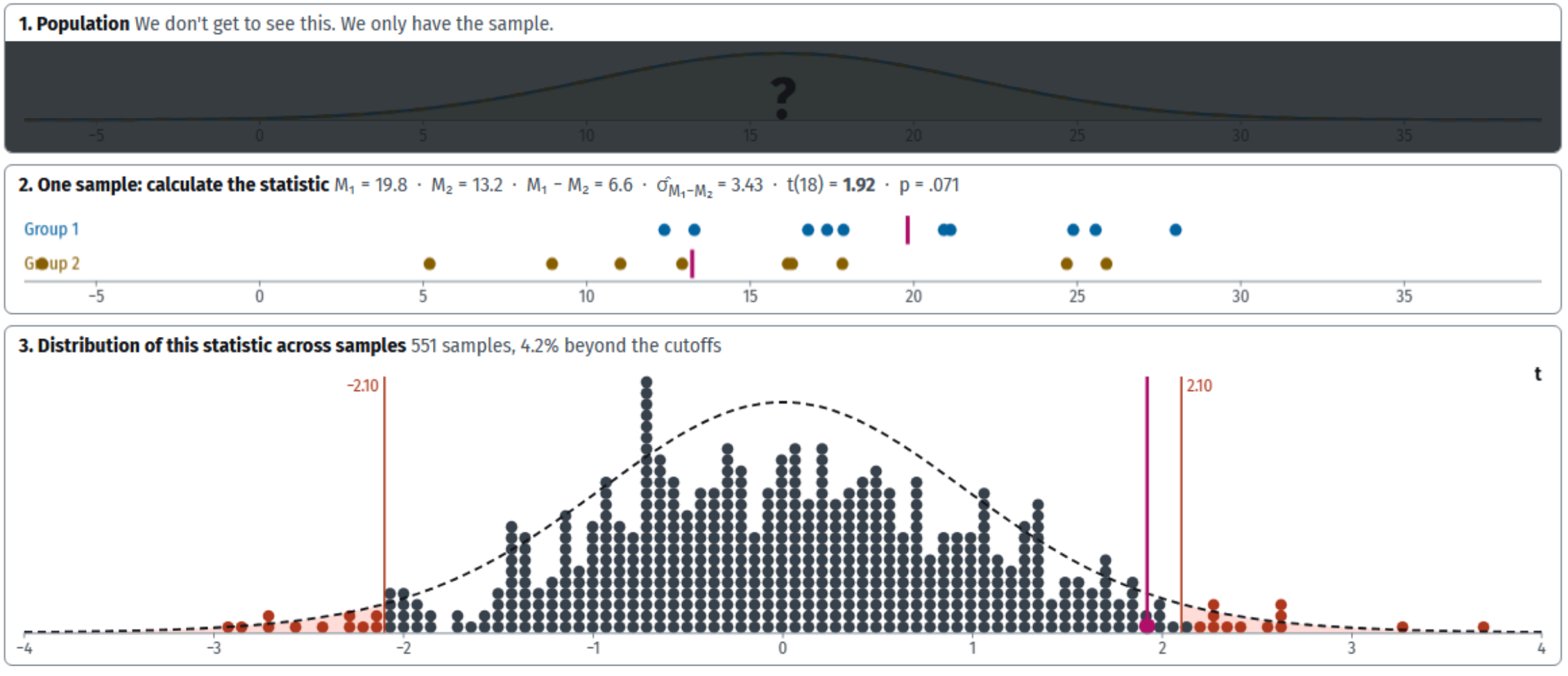
When the groups differ, the pile moves



Simulator after Kristoffer Magnusson, rpsychologist.com.

The two-sample t

In real research, panel 1 goes dark again



Simulator after Kristoffer Magnusson, rpsychologist.com.

The fraction keeps its shape

$$t = \frac{M - \mu}{\hat{\sigma}_M} \qquad \text{becomes} \qquad t = \frac{(M_1 - M_2) - (\mu_1 - \mu_2)}{\hat{\sigma}_{M_1 - M_2}}$$

Top: the gap we saw, minus the gap H_0 claims. Under H_0, that second part is 0

Bottom: how far a gap typically falls. We have to estimate it

In the book: $\hat{\sigma}_{M_1-M_2}$ is written $s_{M_1-M_2}$

Pool the two estimates of σ²

$$\hat{\sigma}_p^2 = \frac{SS_1 + SS_2}{df_1 + df_2}$$

We assume one σ², so we pool: add up the squared deviations and the df

A bigger group holds more information, so it counts for more

The p is for "pooled". With equal n, it's the plain average of the two variances

In the book: $\hat{\sigma}^2_p$ is written s^2_p

Then the bottom of the fraction, and df

$$\hat{\sigma}_{M_1 - M_2} = \sqrt{\frac{\hat{\sigma}_p^2}{n_1} + \frac{\hat{\sigma}_p^2}{n_2}}$$

$$df = (n_1 - 1) + (n_2 - 1)$$

Same formula as before, with the pooled estimate in place of σ^2

Each sample spends one df on its own mean

Three moves: pool the variance, find the standard error, compute t

Side by side with Topic 8

	One sample	Two samples
Estimate	M	$M_1 - M_2$
Under H_0	μ	$\mu_1 - \mu_2 = 0$
Variance	$\hat{\sigma}^2 = SS/df$	$\hat{\sigma}_p^2 = (SS_1 + SS_2)/(df_1 + df_2)$
Standard error	$\hat{\sigma}_M = \sqrt{\hat{\sigma}^2/n}$	$\hat{\sigma}_{M_1 - M_2} = \sqrt{\hat{\sigma}_p^2/n_1 + \hat{\sigma}_p^2/n_2}$
df	$n - 1$	$(n_1 - 1) + (n_2 - 1)$

Every row is the same idea, built for two groups

Try it: 9 fifth graders

Mindspark group: n = 3, M = 20.67, SS = 12.67. Control group: n = 6, M = 16.33, SS = 239.33

Find the pooled variance $\hat{\sigma}^2_p$, then the standard error $\hat{\sigma}_{M_1 - M_2}$

Is $\hat{\sigma}^2_p$ nearer 6.33 or 47.87, the two groups' own estimates? Why? Work alone, compare with a neighbor, then we vote

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

Try it: Answer

$\hat{\sigma}^2_p = (12.67 + 239.33) / (2 + 5) = \mathbf{36.00}$

$\hat{\sigma}_{M_1 - M_2} = \sqrt{(36.00 / 3 + 36.00 / 6)} = \mathbf{4.24}$

It's nearer 47.87, the control group's estimate. That group has more students, so it counts for more

A plain average would have given 27.10. Pooling doesn't treat 3 students like 6

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

Mindspark's test

Steps 1 and 2: set up the test

At endline, do math scores differ across the two groups?

Step	In words	In symbols
1. State H_0	The two population means don't differ	$\mu_{Mindspark} - \mu_{control} = 0$
2. Set the criteria	$\alpha = 0.05$, two tails, and σ is unknown	$df = (266 - 1) + (273 - 1) = 537$
	Reject H_0 if t is beyond the cutoffs	$t(537)$ beyond ± 1.964

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

Step 3: three moves

	n	M	SS
Mindspark	266	18.68	9107.5
Control	273	15.99	9060.0

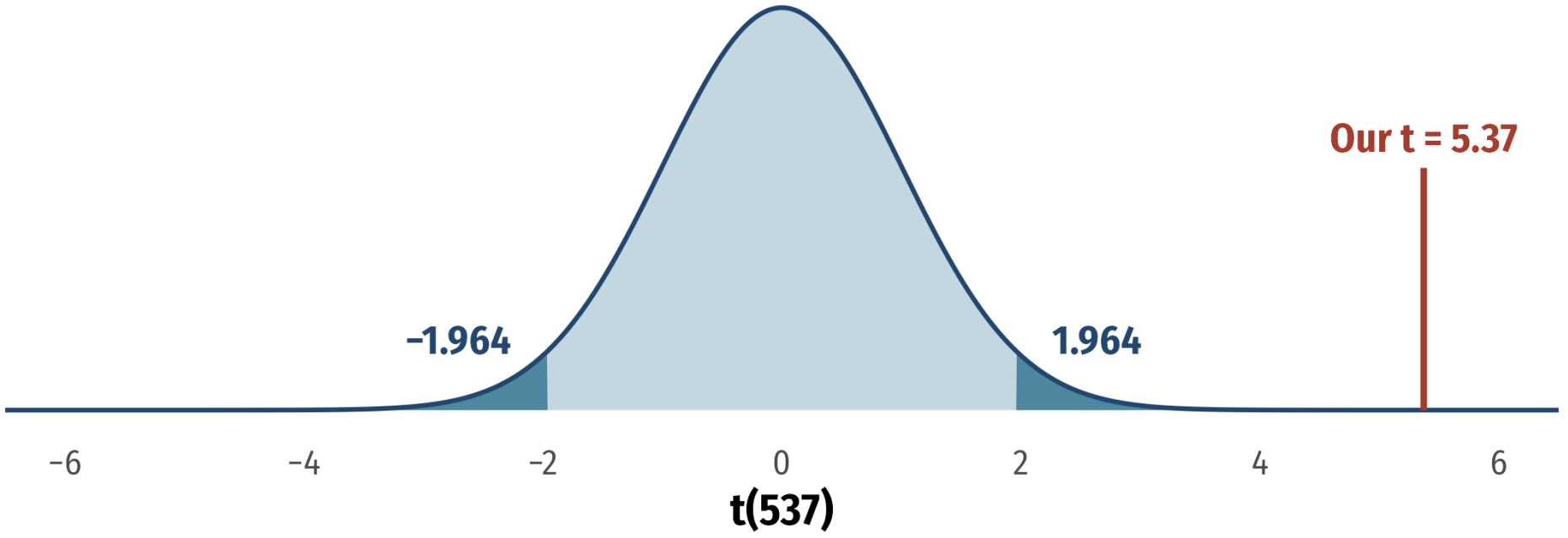
Pool: $\hat{\sigma}^2_p = (9107.5 + 9060.0) / 537 = 33.83$

Standard error: $\hat{\sigma}_{M_1 - M_2} = \sqrt{(33.83 / 266 + 33.83 / 273)} = 0.50$

Compute t: (18.68 − 15.99) / 0.50 = 5.37

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

Step 4: our t is far past the cutoff



5.37 is beyond 1.964, so we reject H_0. p is below 0.001

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

How big is the gap? g = 0.46

$$d = \frac{\mu_1 - \mu_2}{\sigma} \qquad \text{is estimated by} \qquad g = \frac{M_1 - M_2}{\hat{\sigma}_p}$$

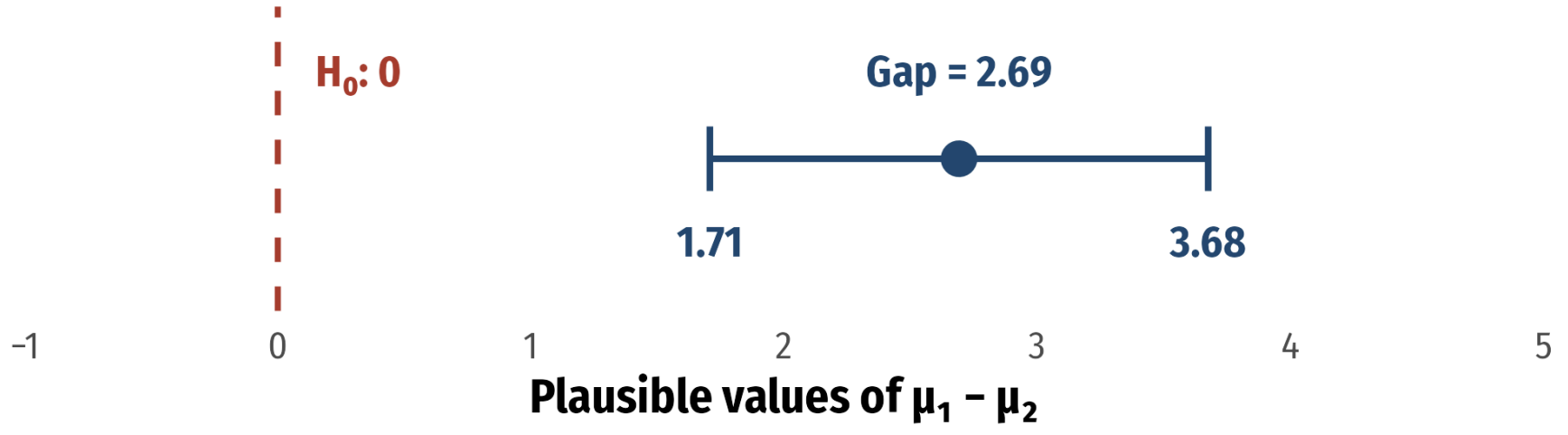
$\hat{\sigma}_p = \sqrt{33.83} = 5.82$, so $g = 2.69 / 5.82 =$ **0.46**

The groups sit about half an SD apart. Remember Topic 7: that's large for education

In the book: g is called the estimated Cohen's d

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

The interval for a gap



Estimate ± margin of error: 2.69 ± 1.964 × 0.501 = [1.71, 3.68]

The interval leaves out 0, so it agrees with the test: reject H_0

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

How we report it

In words

At endline, Mindspark students scored 2.69 points higher in math than control students. The difference is statistically significant at the 0.05 level, t(537) = 5.37, g = 0.46, with a 95% confidence interval of [1.71, 3.68]

The test, the effect size and the interval: all three, as in Topic 8

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

Try it: One Mindspark center

Students at the Govindpuri center. Mindspark: n = 34, M = 19.50, SS = 1408.5. Control: n = 35, M = 16.89, SS = 1455.5

H_0: $\mu_1 - \mu_2 = 0$. $\alpha = 0.05$, two tails. The cutoff is t(67) = ±1.996

Find t and g, then decide. Work alone, compare with a neighbor, then we vote

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

Try it: Answer

$\hat{\sigma}^2_p = (1408.5 + 1455.5) / 67 = \mathbf{42.75}$

$\hat{\sigma}_{M_1-M_2} = \sqrt{(42.75 / 34 + 42.75 / 35)} = \mathbf{1.57}$

$t = 2.61 / 1.57 = \mathbf{1.66}$. That's inside ± 1.996: **maintain H_0**

$g = 2.61 / \sqrt{42.75} = \mathbf{0.40}$

A similar g to the whole study's 0.46, and a different decision. With 69 students, we can't tell

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

Before Tuesday

Friday, February 19: Lab 7, two samples. Bring a calculator and a laptop

Sunday, February 21 at 6 PM: Mastery and Problem Set for Chapter 10

Read Chapter 11 for Tuesday. What if the same people take the test twice?

Session 2: Tuesday, February 23

Thursday in one line: a gap between two groups gets its own standard error, and a pooled $\hat{\sigma}^2$

One group, two scores

Why measure twice?

Who gets the credit?

One group, two scores

Some studies test people twice

Two samples

Different people in each group

One score per person

Compare two groups: Thursday

Paired scores

The same people, before and after

Two scores per person

Compare each person with themselves: today

In the book: independent measures and repeated measures

Two scores become one: the gain

Student	Before (baseline)	After (endline)	Gain = after − before
Student 1	13	18	5
Student 2	9	15	6
Student 3	12	16	4
Student 4	8	14	6
Student 5	7	20	13

Now each student has one number. We've seen what to do with one column of numbers

Data: Muralidharan, Singh, and Ganimian (2019). Five Mindspark students, math.

A paired test is a one-sample t on gains

$$t = \frac{M - \mu}{\hat{\sigma}_M} \qquad \text{where} \qquad \hat{\sigma}_M = \frac{\hat{\sigma}}{\sqrt{n}}$$

There's no new formula. This is Topic 8's, and now X is a person's gain

M is the mean gain, and n is the number of people. H_0 says $\mu = 0$, so df = n − 1

H_0: on average, nothing changes between before and after

In the book: a gain is D, and our M and $\hat{\sigma}_M$ are M_D and s_{M_D}

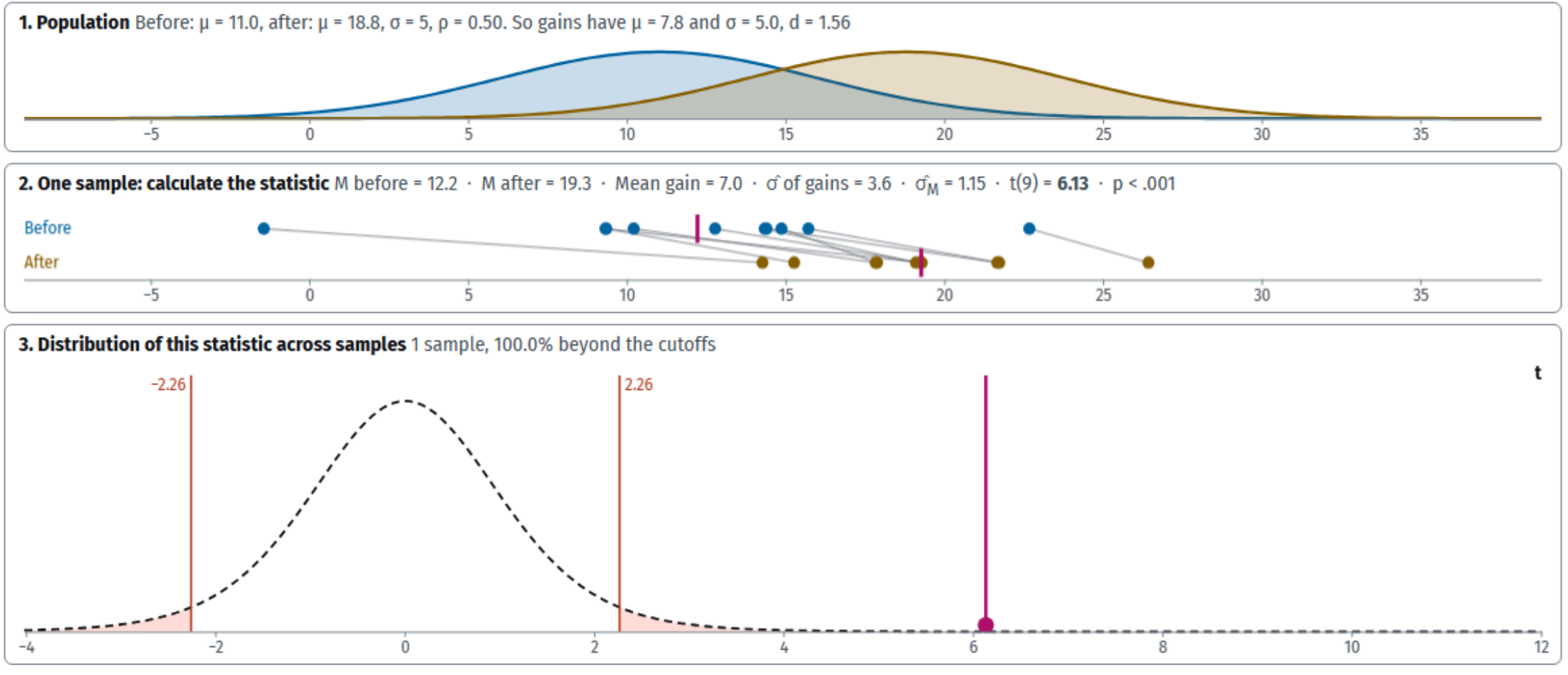
All 266 Mindspark students

Step	Mindspark students' gains
1. State H_0	$\mu = 0$: no change on average
2. Set the criteria	$\alpha = 0.05$, two tails. $t(265)$ beyond ± 1.969
3. Compute t	$M = 7.82$, $\hat{\sigma} = 4.97$, $\hat{\sigma}_M = 4.97 / \sqrt{266} = 0.305$ $t = 7.82 / 0.305 =$ **25.68**
4. Decide	Reject H_0: 25.68 is far beyond 1.969

Mean gain: 7.82 points. g = 1.57, and the 95% interval is [7.22, 8.42]

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

Step 3b: each person gets a line



Simulator after Kristoffer Magnusson, rpsychologist.com.

Try it: Five students' gains

Gains: 5, 6, 4, 6, 13. So M = 6.8 and SS = 50.8

H_0: $\mu = 0$. $\alpha = 0.05$, two tails. The cutoff is $t(4) = \pm 2.776$

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

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

Try it: Answer

$\hat{\sigma} = \sqrt{(50.8 / 4)} = 3.56$

$\hat{\sigma}_M = 3.56 / \sqrt{5} = 1.59$

$t = 6.8 / 1.59 = 4.27$

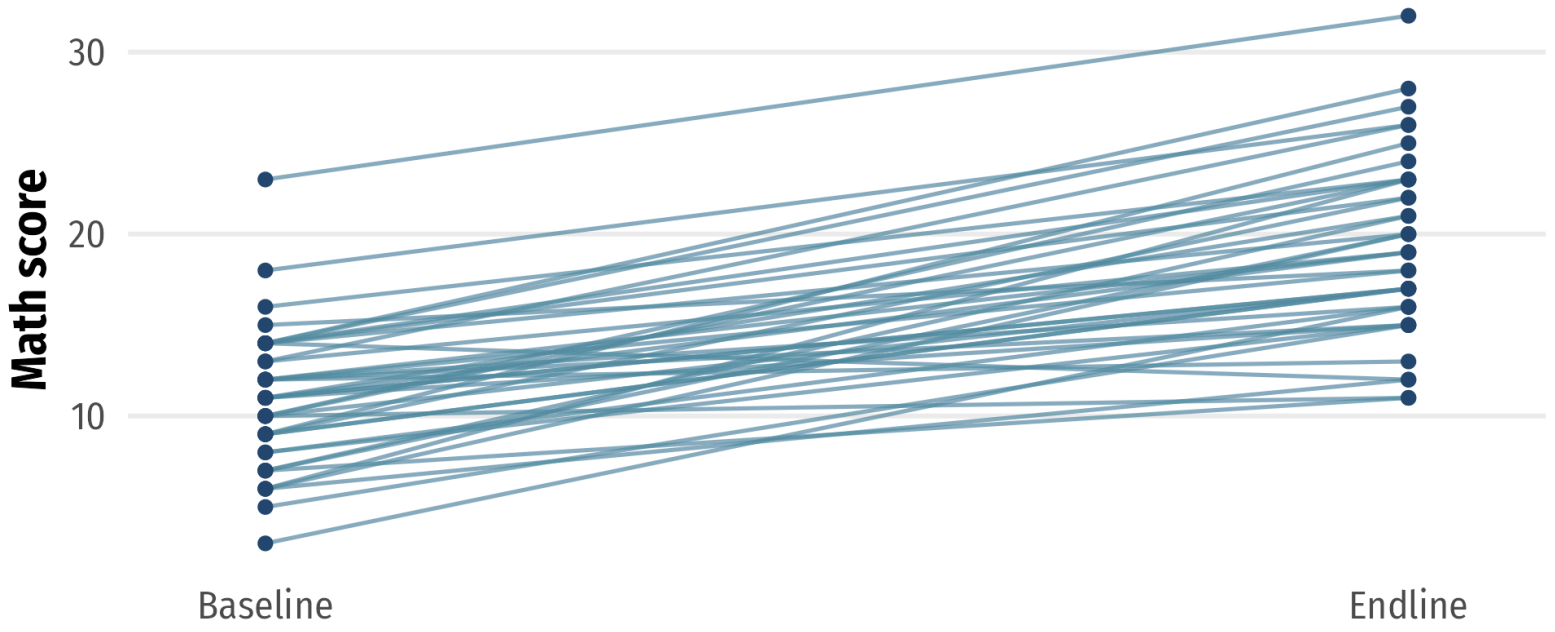
4.27 is beyond 2.776: **reject H_0**

We report: t(4) = 4.27, p = 0.013. Even five students can show a change, if they all move the same way

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

Why measure twice?

Strong students stay fairly strong



A gain takes each student's own starting level out of the picture

Data: Muralidharan, Singh, and Ganimian (2019). 40 random Mindspark students.

Same students, two ways to test

	Standard error	t
Pretending they're two separate groups	0.429	t(530) = 18.23
Using each student's gain	0.305	t(265) = 25.68

Same 266 students, same mean gain of 7.82. Pairing cuts the standard error by 29%

The first row is wrong here: the two columns of scores aren't independent

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

ρ sets the spread of the gains

How tightly before and after move together	ρ	SD of gains
Not at all	0.0	7.07
Moderately	0.5	5.00
Very tightly	0.9	2.24

Each score here has σ = 5. The more tightly they move, the steadier the gains

A smaller SD of gains means a smaller $\hat{\sigma}_M$, and so a bigger t

Predict it: Tighten the scores

Keep the mean gain the same. Raise ρ from 0.5 to 0.9, so before and after move together much more tightly

What happens to the pile of t-scores in panel 3?

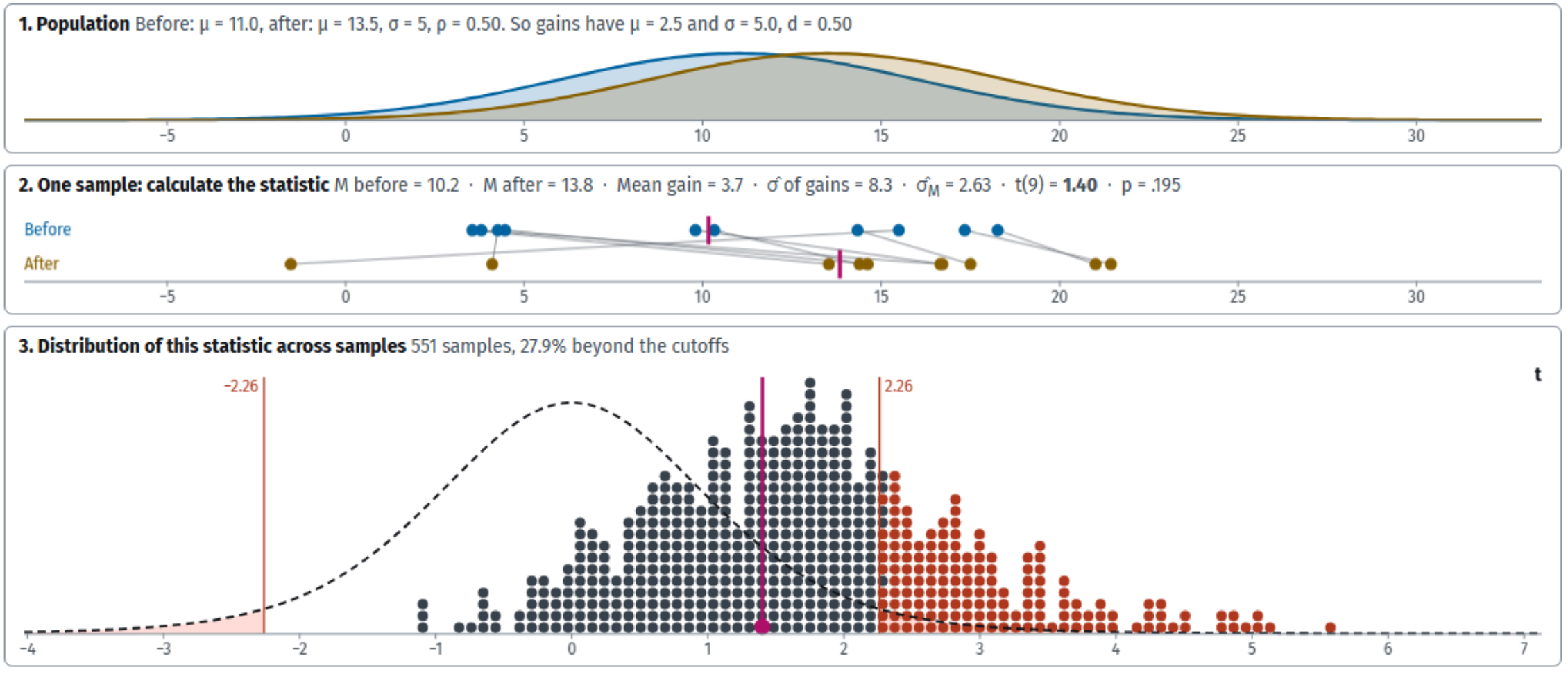
A It moves away from 0

B It moves toward 0

C It stays where it is

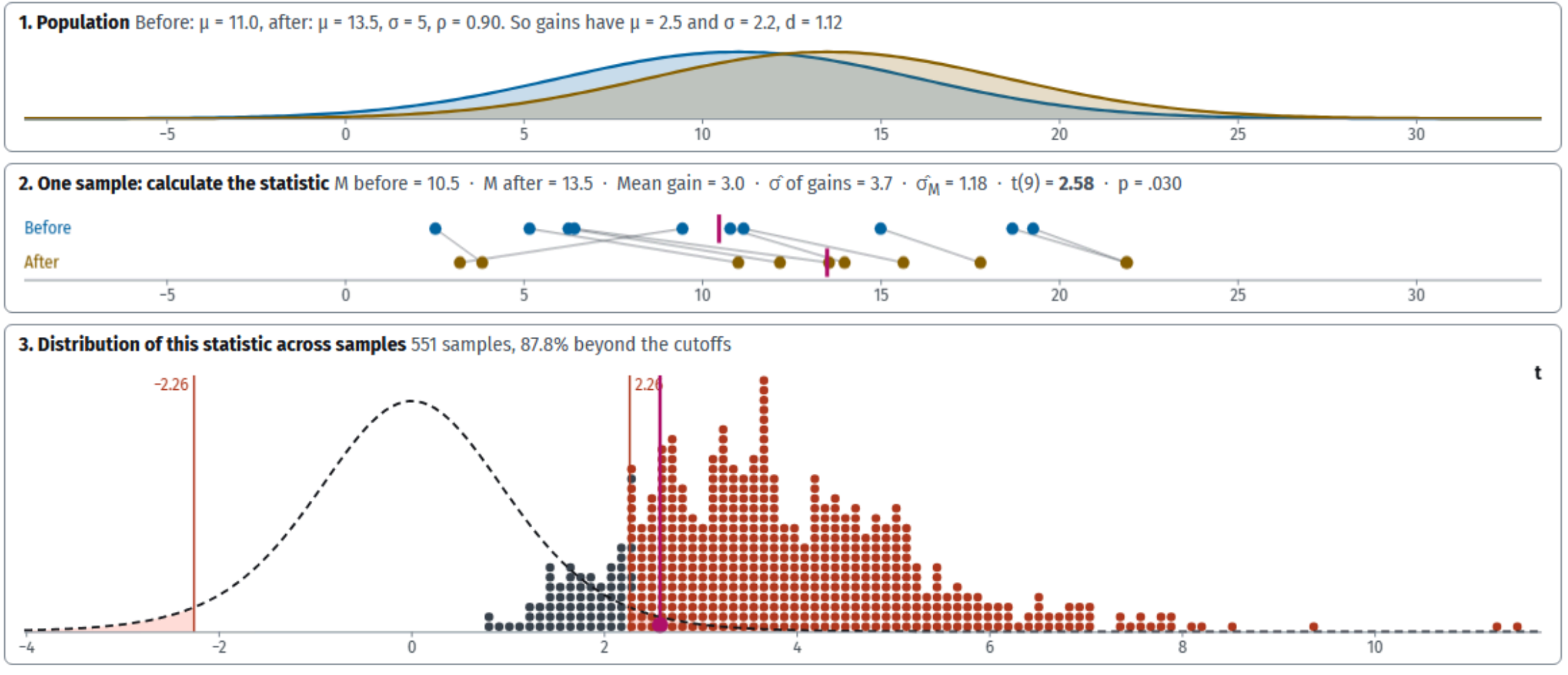
D It depends on the mean before

Predict it: With ρ = 0.5



Simulator after Kristoffer Magnusson, rpsychologist.com.

Predict it: With ρ = 0.9, a bigger t



Simulator after Kristoffer Magnusson, rpsychologist.com.

Pairing has a price

What we gain

- Each person is their own comparison
- A smaller standard error
- Fewer people for the same power

What we risk

- Time passes between the two scores
- Anything that changed gets mixed into the gain
- The test can't say what caused it

Who gets the credit?

Did Mindspark add 7.82 points?

Students gained 7.82 points, t(265) = 25.68

The test is right: their scores really rose. Sampling error can't explain that

Imagine a press release: "Students gain nearly 8 points with Mindspark!"

What else was going on for these students during those months?

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

Mindspark wasn't all that happened

School went on for 4.5 months

Children got older

They had seen a math test like this one before

The endline test wasn't the same test

A paired test can't separate any of this from the program

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

Predict it: Students with no program

Respond at **PollEv.com/adb**

Control students lost the lottery, and had no Mindspark during those 4.5 months. On average, how much did their math scores rise?

A About 0 points

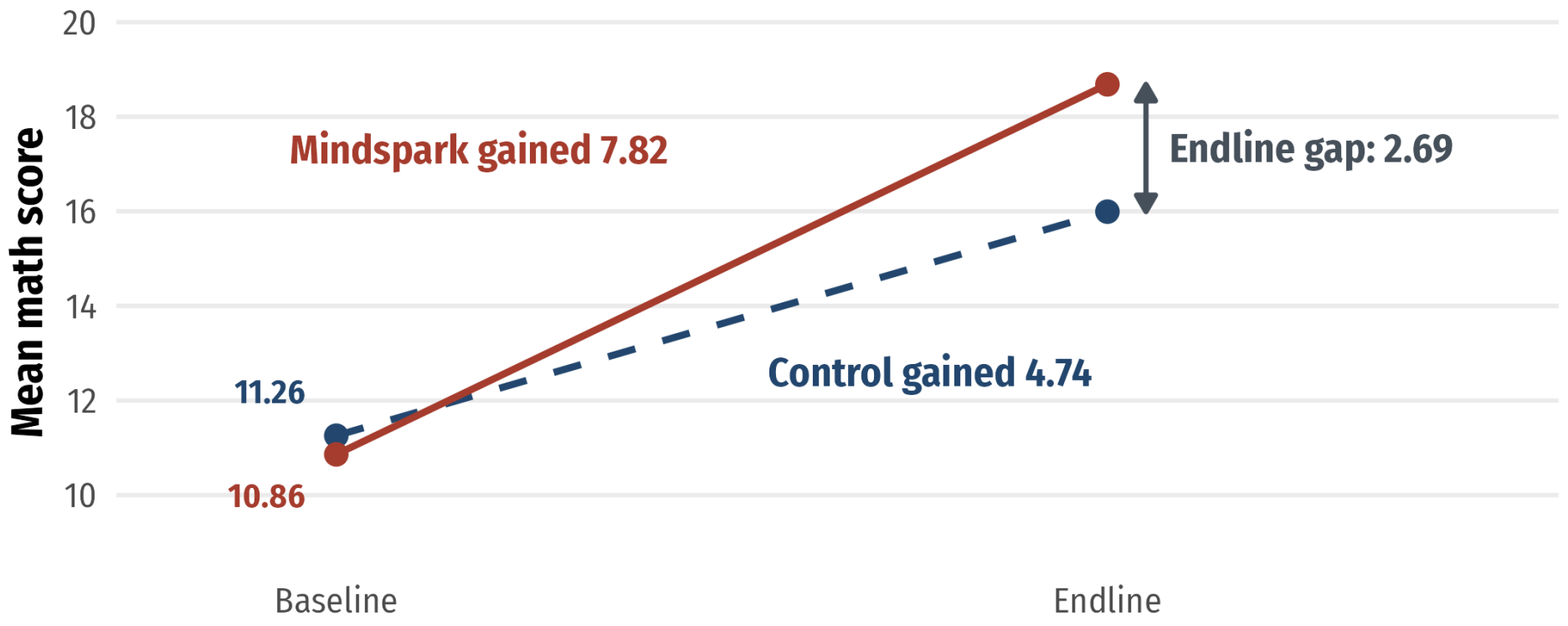
B About 2 points

C About 5 points

D About 8 points, like the Mindspark group

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

Both groups rose, one rose more



Data: Muralidharan, Singh, and Ganimian (2019). The 539 students with both scores.

A paired test alone over-credits

What we'd conclude	Points	Where it comes from
Paired test, Mindspark students alone	7.82	Gain: t(265) = 25.68
Paired test, control students alone	4.74	Gain: t(272) = 15.48
Two-sample test at endline	**2.69**	Gap: t(537) = 5.37

7.82 / 2.69 = 2.9: nearly three times the program's effect

The paired test isn't wrong. It asks if scores changed, not why

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

Why doesn't 7.82 – 4.74 give 2.69?

Subtracting gains gives 3.09, a two-sample t on gains: t(537) = 7.15

It differs from 2.69 by 0.40: the gap at baseline, which favored the control group by chance

Both are fair estimates, because both compare with the control group

The endline gap ignores baseline. Gains use it, so their standard error is smaller: 0.43, not 0.50

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

This is why control groups exist

What would have happened without the program?

We can't watch the same students both with and without Mindspark

The lottery gave us the next best thing: similar students, who lived through the same months

The control group's rise is what would have happened anyway. Whatever goes beyond it is the program

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

Play it: Name that test

1 One-sample t · 2 Paired t · 3 Two-sample t. Hold up your fingers

A 40 students take a reading test in September and again in June

B Tutored students and untutored students each take one final exam

C 25 teachers' scores are compared with a national mean of 100

D Each child is matched with a similar child, and one of each pair gets the program

Play it: Answers

Clue	Test	The giveaway
A	Paired t	Same students, two scores. But no control group
B	Two-sample t	Different students, one score each
C	One-sample t	One mean against a known value
D	Paired t	Each pair gives one difference

Ask two things: How many scores per person? Is there a comparison group?

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: next	F	Topic 10	

Same fraction every time: what we saw, over what sampling error alone would give

Where are we?

Description

Summarize data

Visualize data

Prediction

Did scores change?

Explanation

Did the program cause it?

Needs a comparison

What you can do now

Run a two-sample t-test: pool, find the standard error, compute t

Report it with Hedges' g and a confidence interval

Turn before and after scores into gains, and run a paired t-test

Say why pairing shrinks the standard error

Explain why a gain isn't an effect, and what a control group adds

Next steps

Thursday, February 25: Topic 10

ANOVA. Read sections 12-1 to 12-4 and 12-6, and the start of 12-5

Friday, February 26: Lab 8

A 15-minute individual check on two samples, then gains and ANOVA. Bring a calculator and a laptop

Sunday, February 28, 6 PM: Chapters 11 and 12 are due in MindTap