Hypothesis testing

Topic 7 · Thursday, January 28, and Tuesday, February 2

EDUC 15: Statistics for Education Research
UCI School of Education

Where are we?

Description

Summarize data

Visualize data

Prediction

From a sample to a population

We are here



Peter the Anteater at a UC Irvine basketball game.

Explanation

The question we keep asking

If what I think about the population is true, how weird is my sample?

Topic 6 gave us the tool: $z = (M - \mu) / \sigma_M$ says how far a mean is from μ

Today's new question: how weird must it be before I change my mind?

Plan for today

Two steps before any data

Observe and decide

When the decision is wrong

Two steps before any data

The four steps of a hypothesis test

Step	What we do	On the blueprint
1. State a null hypothesis, H_0	Say what we assume about the population	Panel 1
2. Set criteria	Say what would be too unexpected if H_0 is true	Panel 3
3. Observe a sample	Ask how unexpected this one sample is	Panel 2
4. Reject or maintain H_0	Decide what to believe about the population	Back to panel 1

These same four steps return in every test this quarter: z, t, F and regression

Step 1: State a null hypothesis

The null hypothesis, H_0, says that in the population there's no change, no difference or no relationship, or it names a specific value

Our example: a reading test has μ = 100 and σ = 15. A school tries a new program with 10 students

$H_0: \mu = 100$

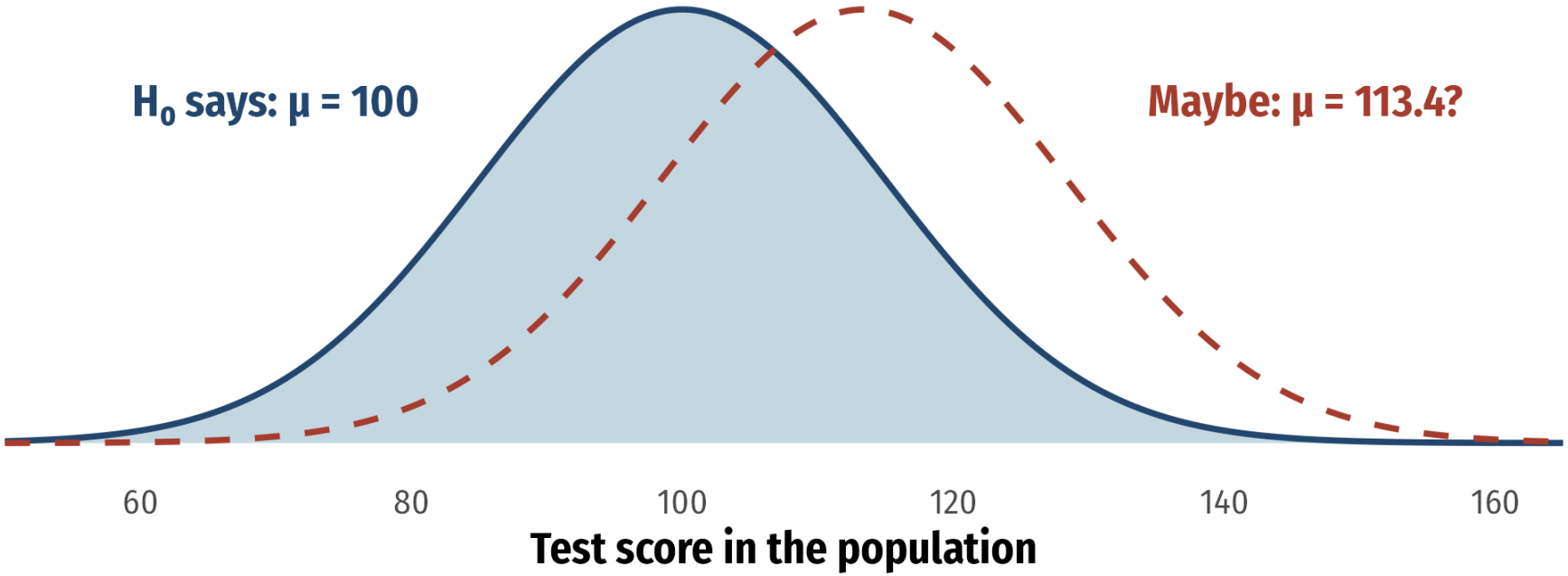
The program changes nothing

$H_1: \mu \neq 100$

The program changes the mean

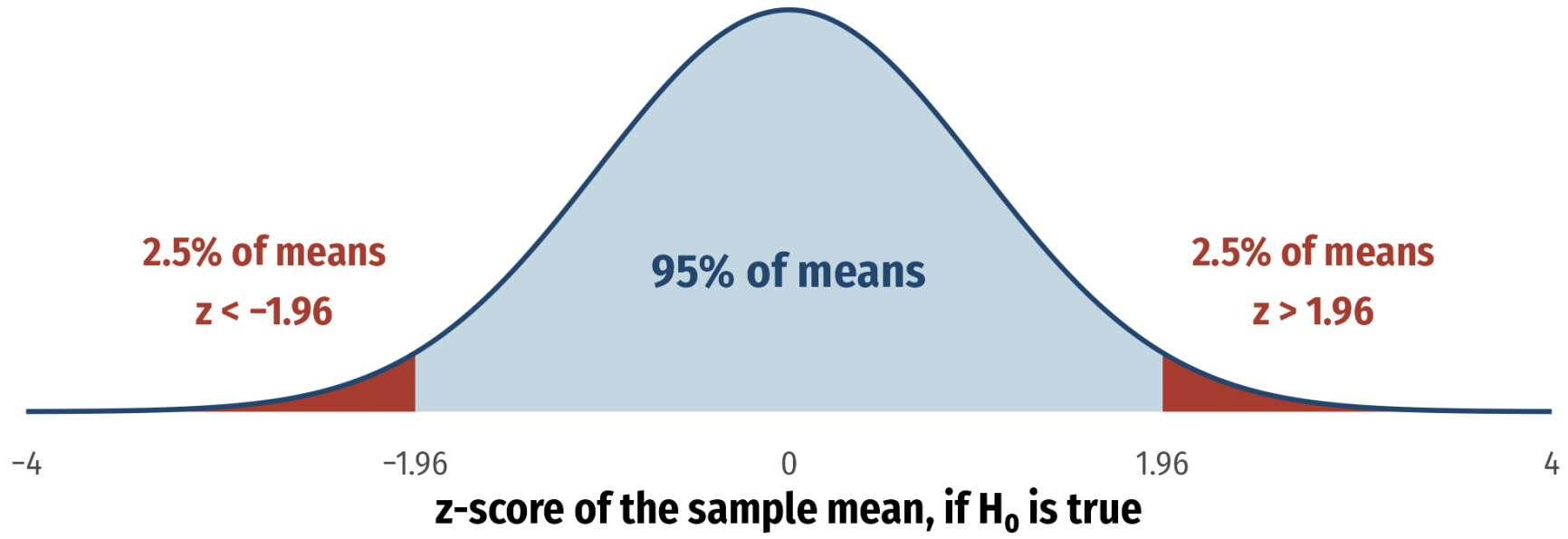
H_1 is the alternative hypothesis. It's everything H_0 rules out

The truth stays hidden



Maybe the program really moved μ to 113.4. We can't see it

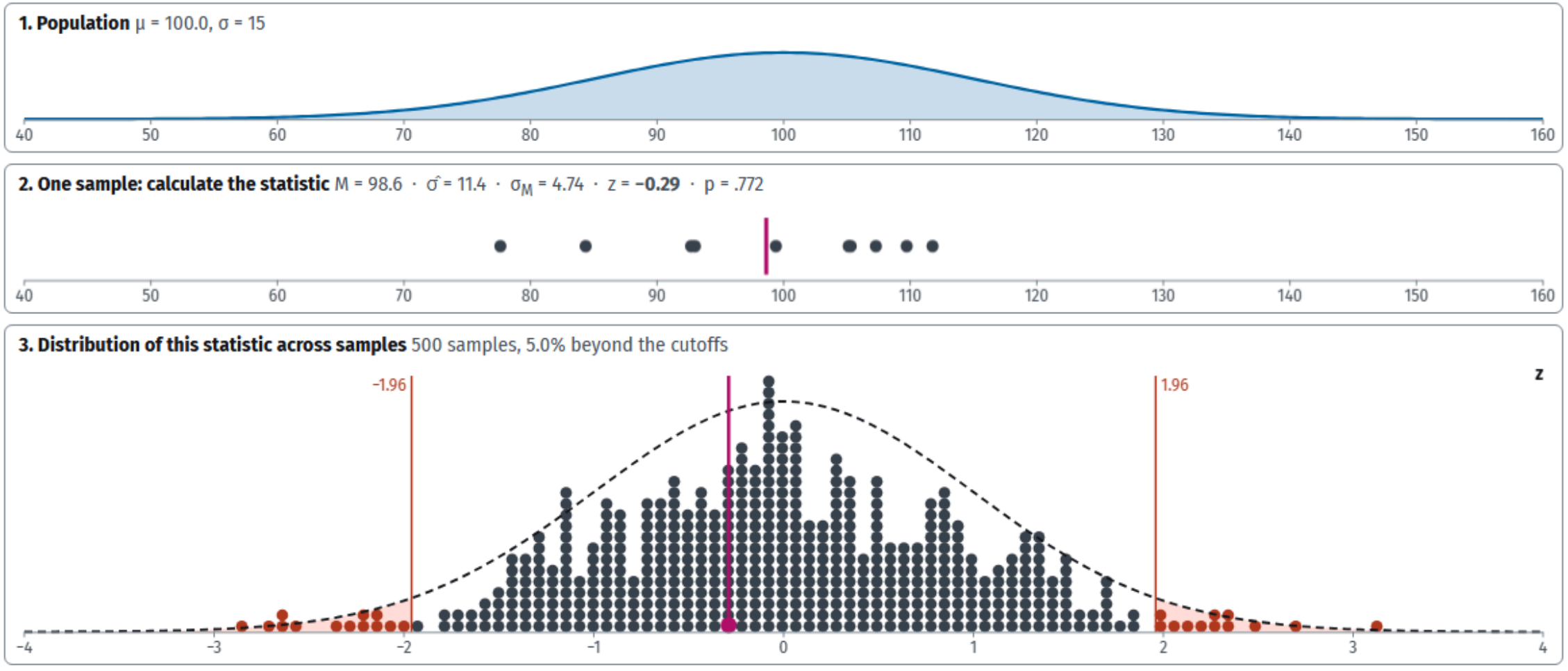
Step 2: Say what's too unexpected



The **alpha level**, α, is our cutoff. The usual choice is α = 0.05

The **critical region** holds the 5% of means least likely if H_0 is true

About 5% land beyond the cutoffs



Simulator after Kristoffer Magnusson, rpsychologist.com.

Predict it: Ten times the sample

Respond at **PollEv.com/adb**

H_0 is still true. With n = 10, about 5% of z-scores landed beyond ±1.96. Now each sample has 100 people. What share lands beyond ±1.96?

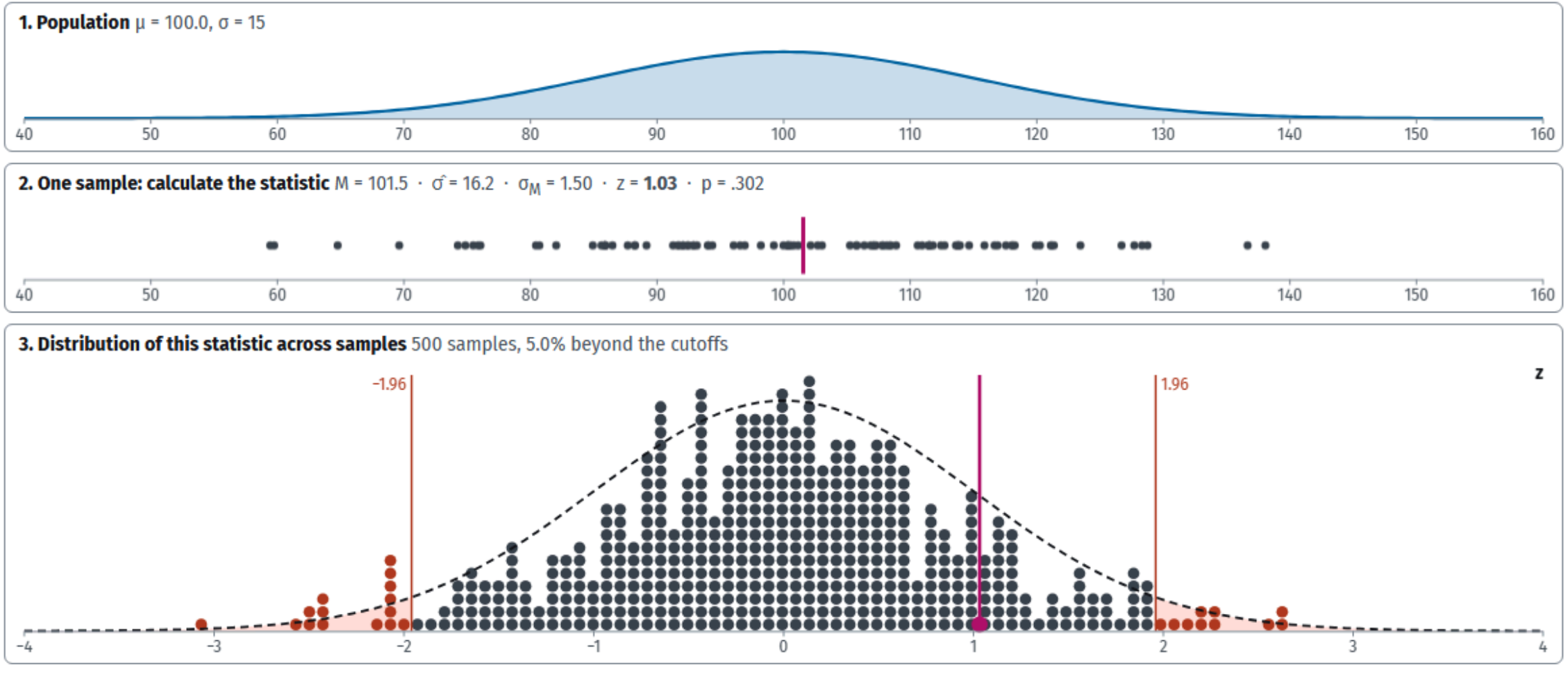
A Far less than 5%: bigger samples are more accurate

B About 5%

C Far more than 5%

D It depends on the value of μ

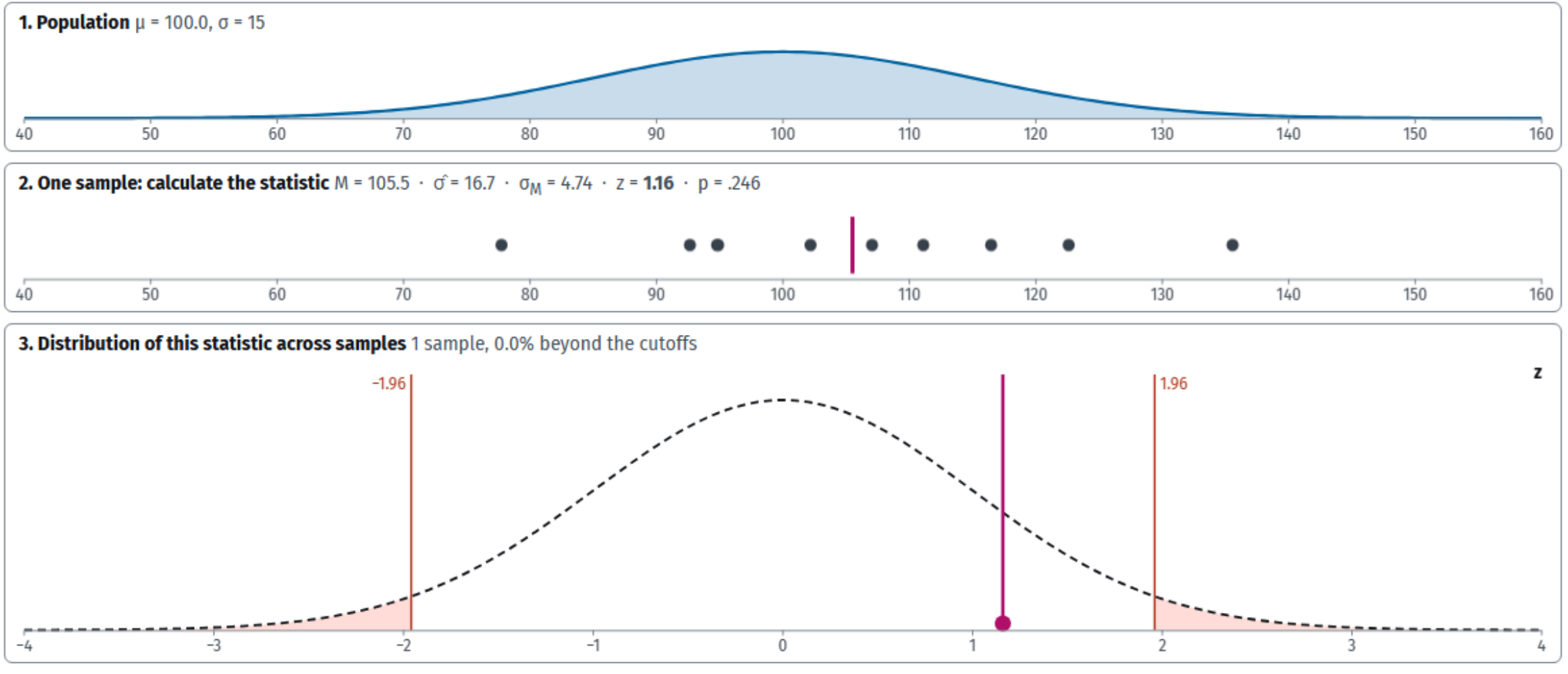
Predict it: Still about 5% at n = 100



Simulator after Kristoffer Magnusson, rpsychologist.com.

Observe and decide

Step 3: Observe one sample



Simulator after Kristoffer Magnusson, rpsychologist.com.

Step 3: How far, in standard errors?

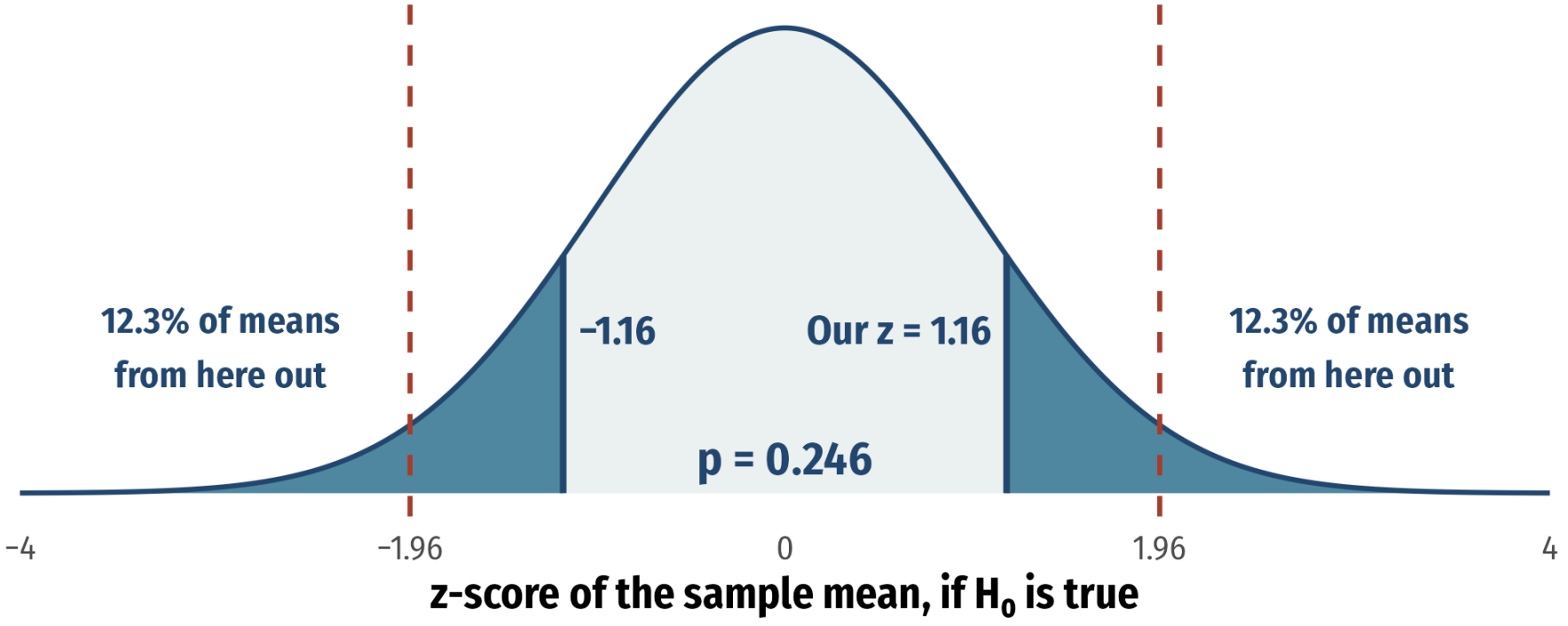
First, the standard error

$$\sigma_M = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{10}} = 4.74$$

Then the test statistic: our mean sits 1.16 standard errors above what H_0 says

$$z = \frac{M - \mu}{\sigma_M} = \frac{105.5 - 100}{4.74} = 1.16$$

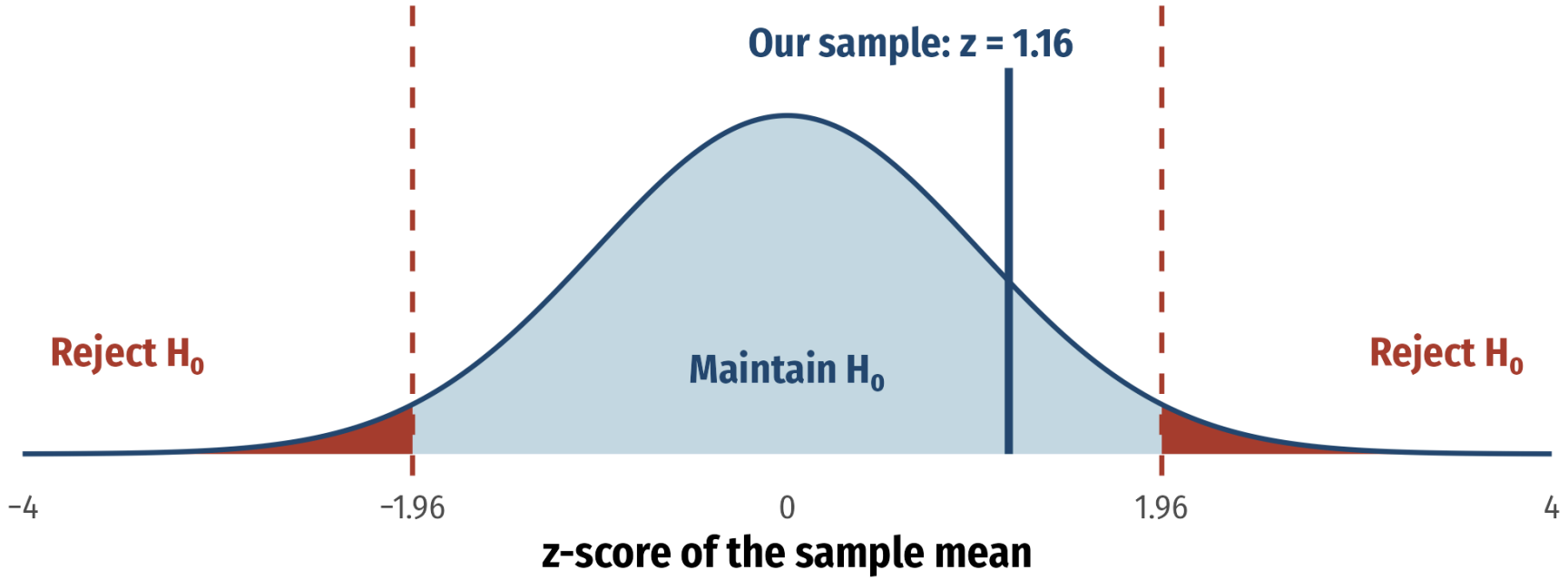
Step 3: The p-value is an area



The **p-value** is the probability of a mean at least this extreme, if H_0 is true

Step 4: p is above α, so we maintain H_0

1. State a hypothesis	2. Set criteria	3. Observe	4. Decide
H_0: μ = 100	α = 0.05	−1.96 < z < 1.96 p = 0.246 > 0.05	Here, maintain H_0



Maintaining H_0 doesn't prove it

We looked for evidence against H_0 and didn't find enough

A sample with M = 105.5 fits μ = 100

It also fits μ = 104, or μ = 108

So we say "maintain H_0", and never "H_0 is true"

In the book: "fail to reject H_0", which means the same

What the p-value doesn't tell you

A common reading	Why it's wrong
"There's a 24.6% chance that H_0 is true"	p assumes H_0 is true, so it can't be the chance that it is
"A small p means a big effect"	p also depends on n. We show this on Tuesday
"p > 0.05 means there's no effect"	The sample may just be too small to show it

p answers one question: if H_0 is true, how unusual is a sample like mine?

The four steps, ready to use

Step	In practice
1. State H_0	H_0: μ = some value, in the population
2. Set criteria	Pick α. Find the critical values, such as $z = \pm 1.96$
3. Observe a sample	Compute the test statistic: $z = (M - \mu) / \sigma_M$. Find p
4. Reject or maintain H_0	Reject if z is in the critical region, which is the same as $p < \alpha$

z and p always agree on the decision. Use whichever is quicker

Try it: The ninth graders

Treat all 619 Mindspark students as a population: at baseline, $\mu = 11.12$ and $\sigma = 3.90$ in math

The 84 ninth graders have $M = 12.39$. If they were a random draw from the 619, would that mean be too unexpected?

Run all four steps, with $\alpha = 0.05$

Work alone for two minutes. Compare with a neighbor. Then we vote: reject or maintain?

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

Try it: Answer

1. State a hypothesis
H_0: $\mu = 11.12$

2. Set criteria
$\alpha = 0.05$, so $z = \pm 1.96$

3. Observe
$z = 2.99$
$p = 0.003 < 0.05$

4. Decide
Here, reject H_0

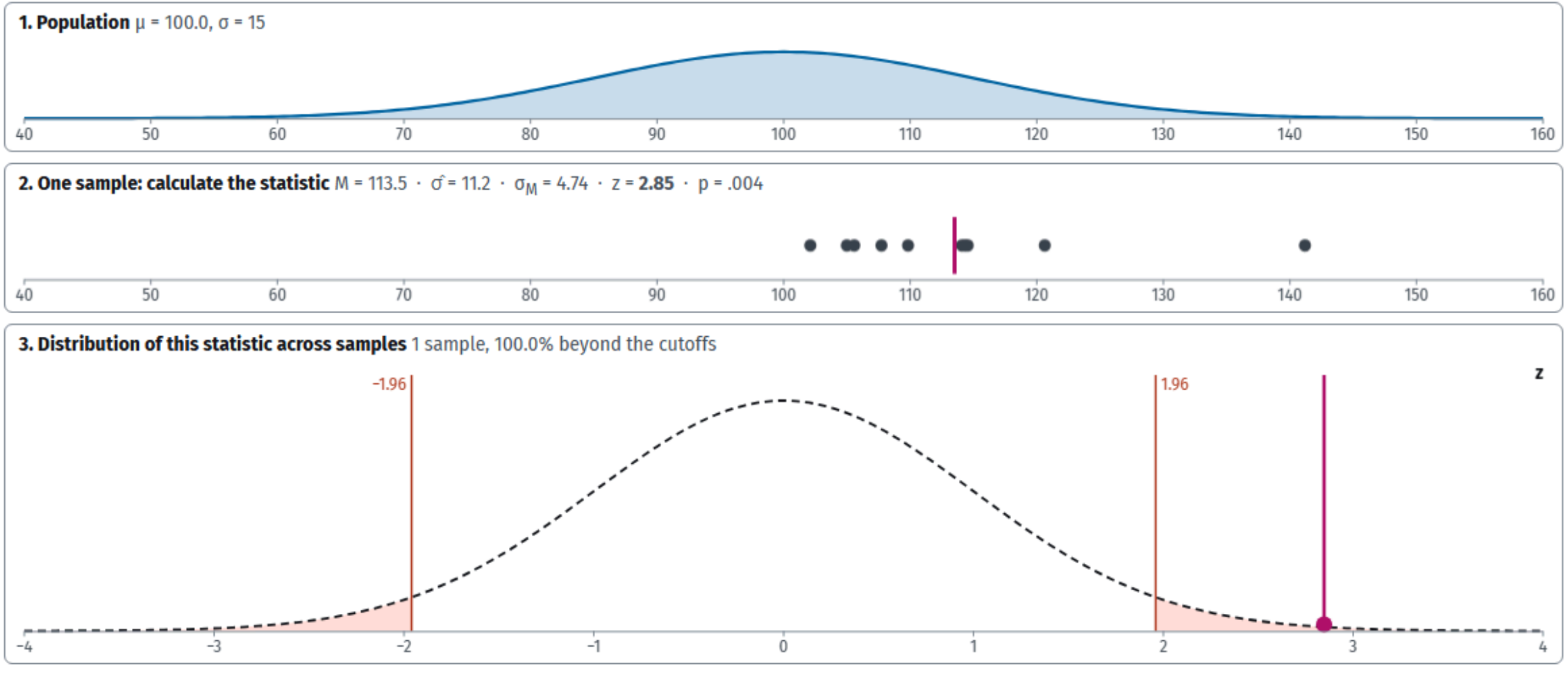
$\sigma_M = 3.90 / \sqrt{84} = 0.43$, so $z = (12.39 - 11.12) / 0.43 \approx 2.99$

They don't look like a random draw from the 619. The test can't say why

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

When the decision is wrong

Same population, a different sample



Simulator after Kristoffer Magnusson, rpsychologist.com.

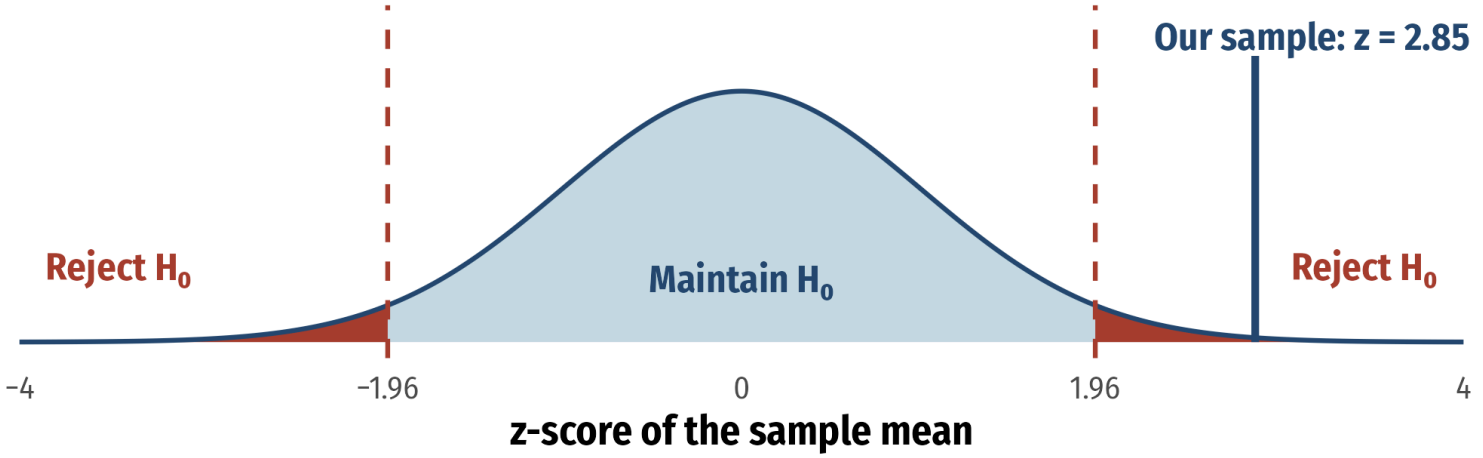
This sample says reject H_0

1. State a hypothesis
$H_0: \mu = 100$

2. Set criteria
$\alpha = 0.05$

3. Observe
$z > 1.96$
$p = 0.004 < 0.05$

4. Decide
Here, reject H_0



We'd write: the mean differs from 100, and it's **statistically significant** at the 5% level ($p < .05$)

But we know H_0 was true here

The simulator set $\mu = 100$, and we still rejected H_0

We would have falsely concluded that $\mu \neq 100$

Nothing went wrong with the sample or the math. This was an unlucky draw

In real research, nobody tells you afterward

Two ways to be wrong

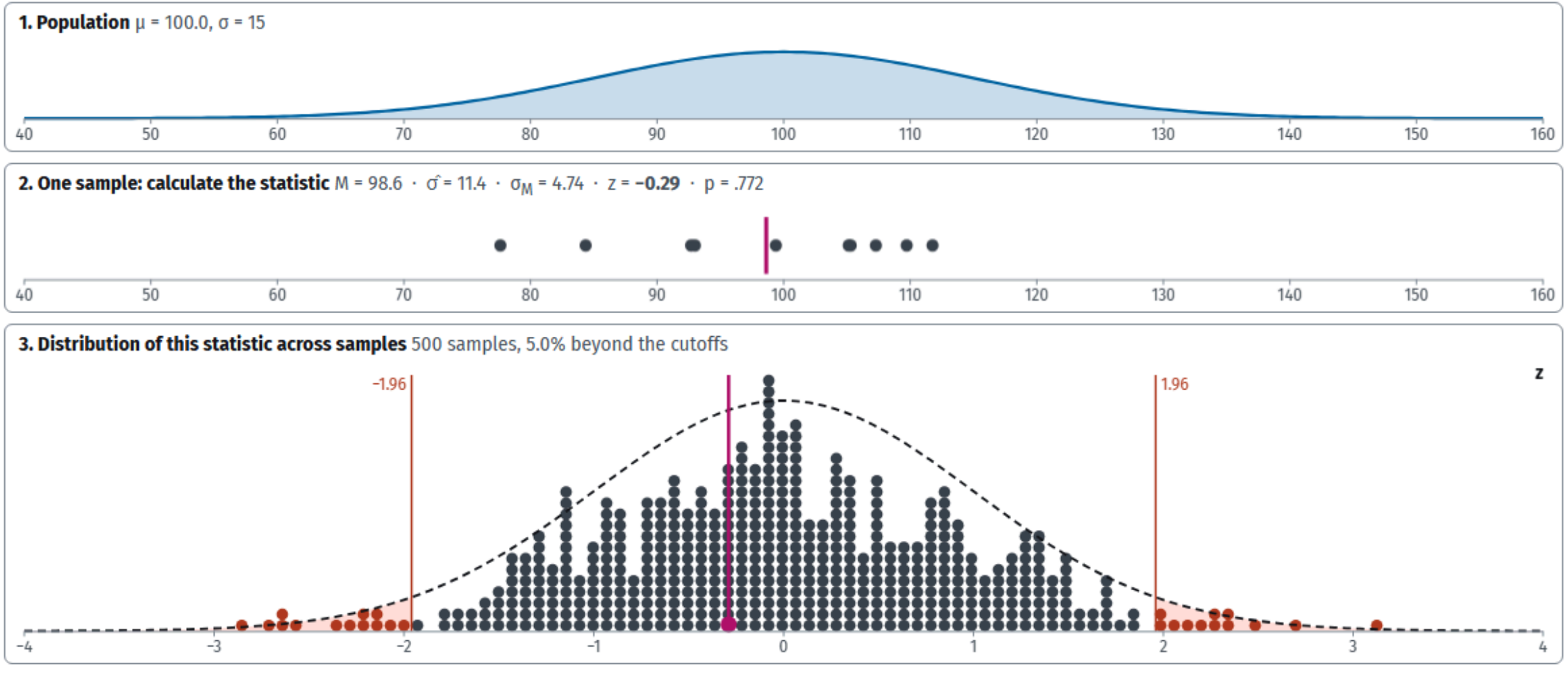
Our decision	If H_0 is true	If H_0 is false
Reject H_0	**Type 1 error**: a false alarm	Correct
Maintain H_0	Correct	**Type 2 error**: a miss

Type 1 error: rejecting H_0 when it's true

Type 2 error: failing to reject H_0 when it's false

In the book and the simulator: Type I and Type II

Every red dot is a false alarm



Simulator after Kristoffer Magnusson, rpsychologist.com.

$\alpha = 0.05$ is a convention, not a law

α	Critical z	Type 1 errors, if H_0 is true
0.10	±1.64	About 1 sample in 10
0.05	±1.96	About 1 sample in 20
0.01	±2.58	About 1 sample in 100

Researchers mostly use 0.05, sometimes 0.01 or 0.10

$p = 0.099$ and $p = 0.101$ are nearly the same evidence, even if $\alpha = 0.10$ splits them. Report the p-value, not just the verdict

Play it: Name that error

100

H_0 is true, and you reject it

200

A program really works, but your small study maintains H_0

300

Lowering α from 0.05 to 0.01 makes this error rarer, and this other one more likely

Answer with a neighbor, as a question. Nothing is scored

Play it: Answers

100

What is a Type 1 error?

A false alarm

200

What is a Type 2 error?

A miss

300

What are Type 1, then Type 2?

A stricter α means fewer alarms, and more misses

Before Tuesday

This Friday: Lab 4

Sampling distributions and the four steps. Bring a calculator and a laptop

Over the weekend

Read sections 8-4 to 8-6. Tuesday we ask: what if H_0 is false?

Mastery and Problem Set for Chapter 7 are due Sunday, January 31 at 6 PM

Session 2: Tuesday, February 2

Thursday in one line: four steps, and two ways to get the decision wrong

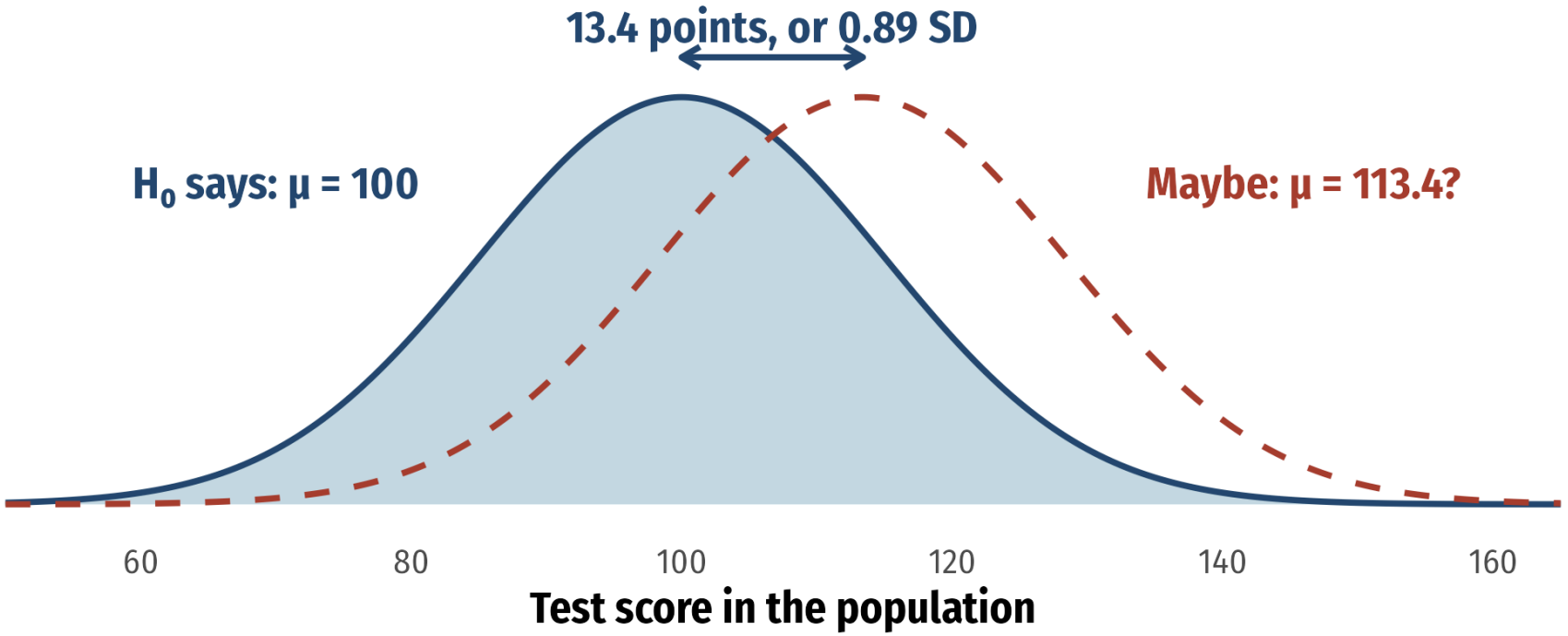
What if H_0 is false?

Power

Planning a study, reading a result

What if H_0 is false?

So far H_0 was true. What if it isn't?



Suppose the program really works, and the true mean is 113.4

Effect size: how big is the gap, in SDs?

$$\text{Cohen's } d = \frac{\mu_1 - \mu_0}{\sigma} = \frac{113.4 - 100}{15} = 0.89$$

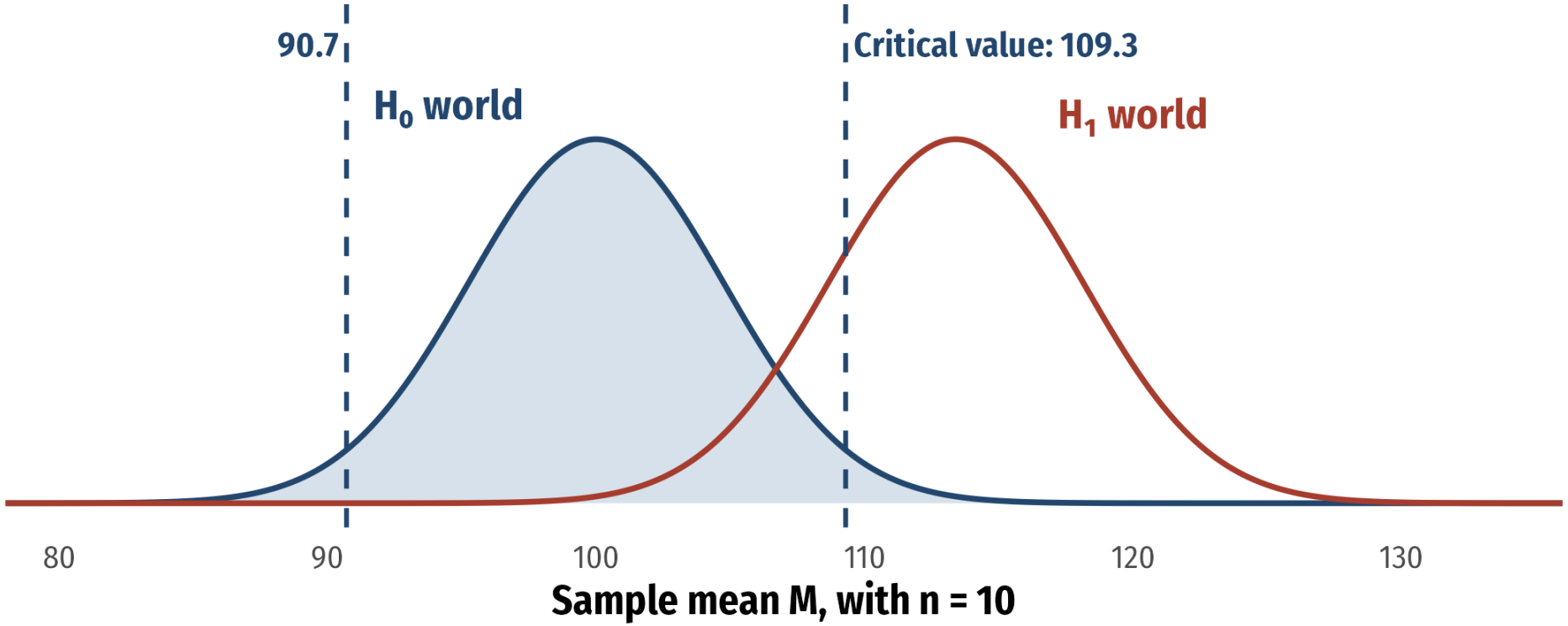
It has no units, so we can compare tests on different scales

n isn't in the formula. It's about the size of the effect, not the sample

d is about the population, so we can only estimate it. Its estimate, Hedges' g, comes in Topic 8

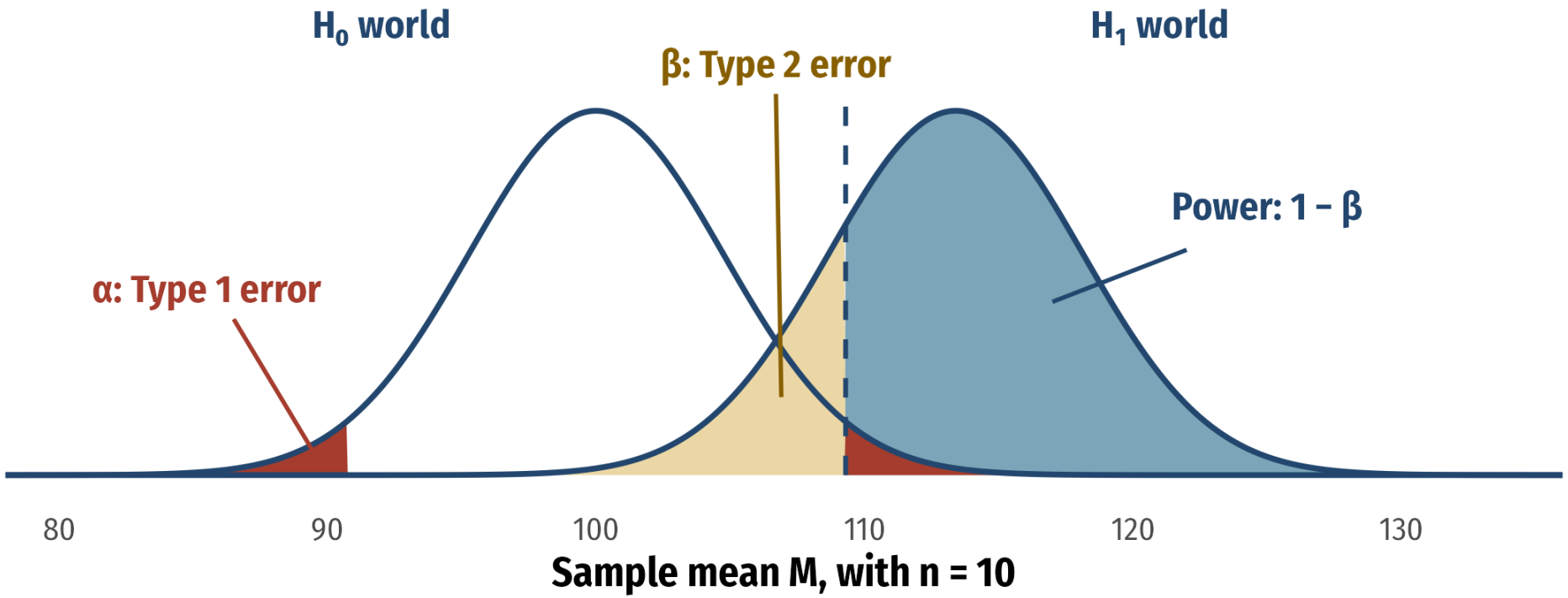
In the book: g is called "estimated d"

Two worlds: where sample means land



Critical values come from the H_0 world. Any mean past them gets "reject"

Each kind of outcome is an area



Power is the chance of rejecting H_0 when it's false. It's the opposite of a miss

Careful: β here is a probability. In regression, it will be a population slope

Predict it: How often do we catch it?

Respond at **PollEv.com/adb**

The program works: $\mu = 113.4$, so $d = 0.89$. We test it with $n = 10$ and $\alpha = 0.05$, two-tailed. In what share of samples do we reject H_0?

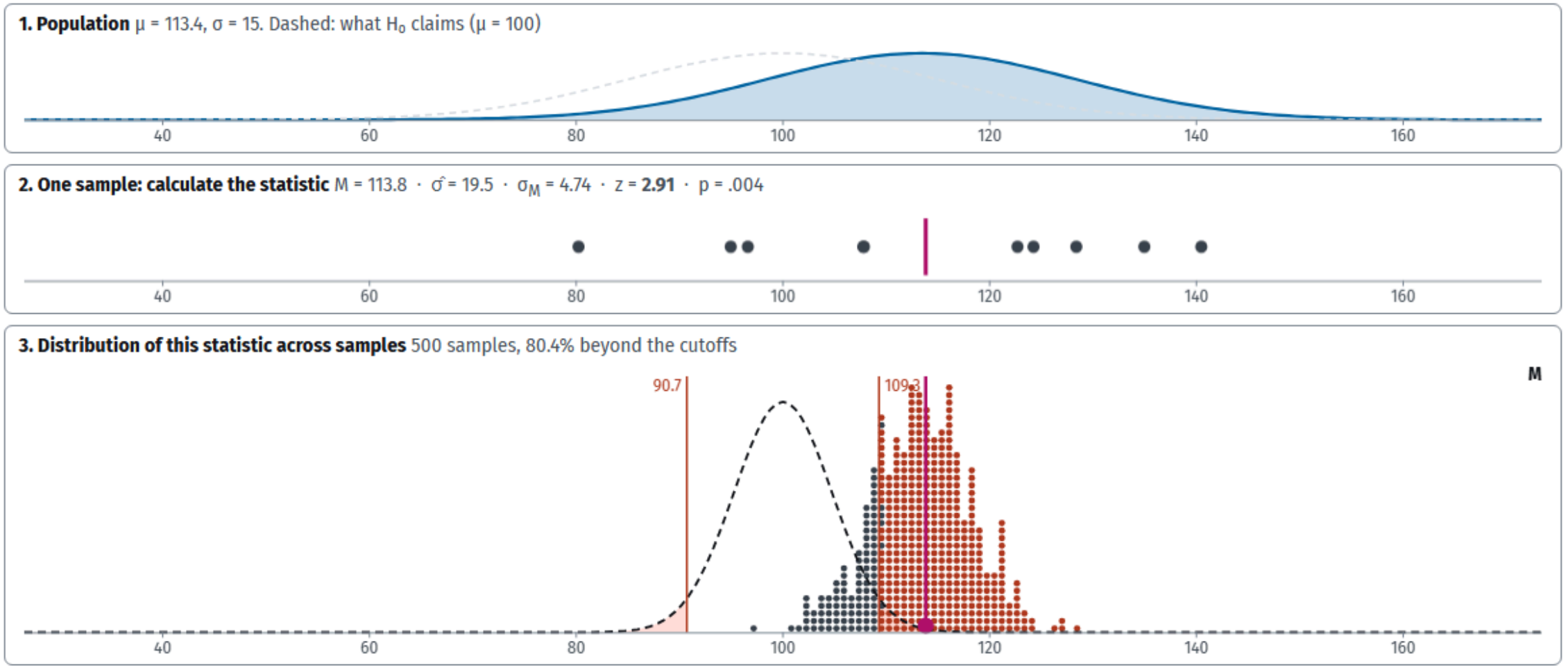
A About 5%

B About 50%

C About 80%

D About 100%

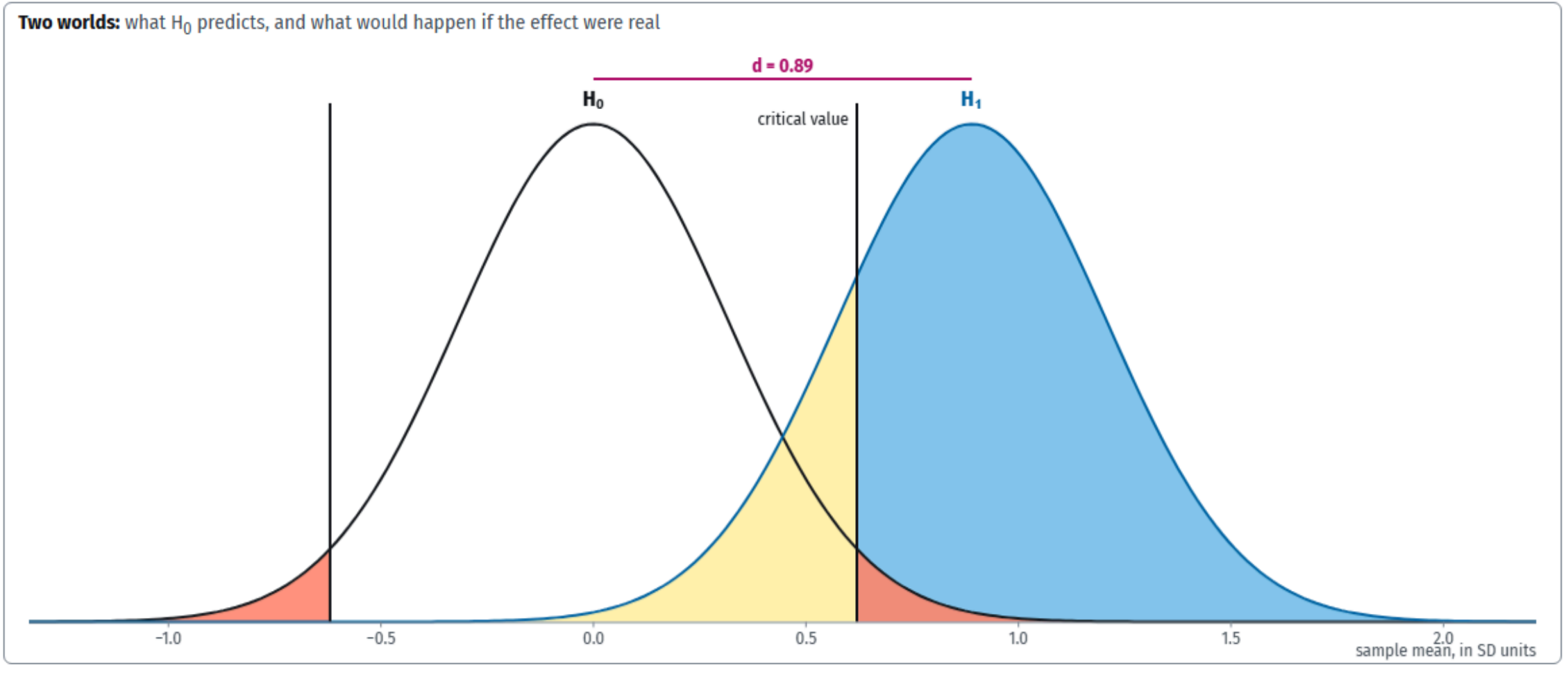
Predict it: About 80% get rejected



Simulator after Kristoffer Magnusson, rpsychologist.com.

Power

Power: catching a real effect



Simulator after Kristoffer Magnusson, rpsychologist.com.

1,000 samples fill in the table

Our decision	H_0 is true ($\mu = 100$)	H_0 is false ($\mu = 113.4$)
Reject H_0	**29** Type 1 errors	**412** correct
Maintain H_0	**471** correct	**88** Type 2 errors

First column: 29 of 500 is 5.8%. That estimates $\alpha = 5\%$

Second column: 412 of 500 is 82.4%. That estimates power = 80%

Simulated in R: 500 samples of 10 from each population.

Four ways to raise power

1. Loosen α: from 5% to 10%

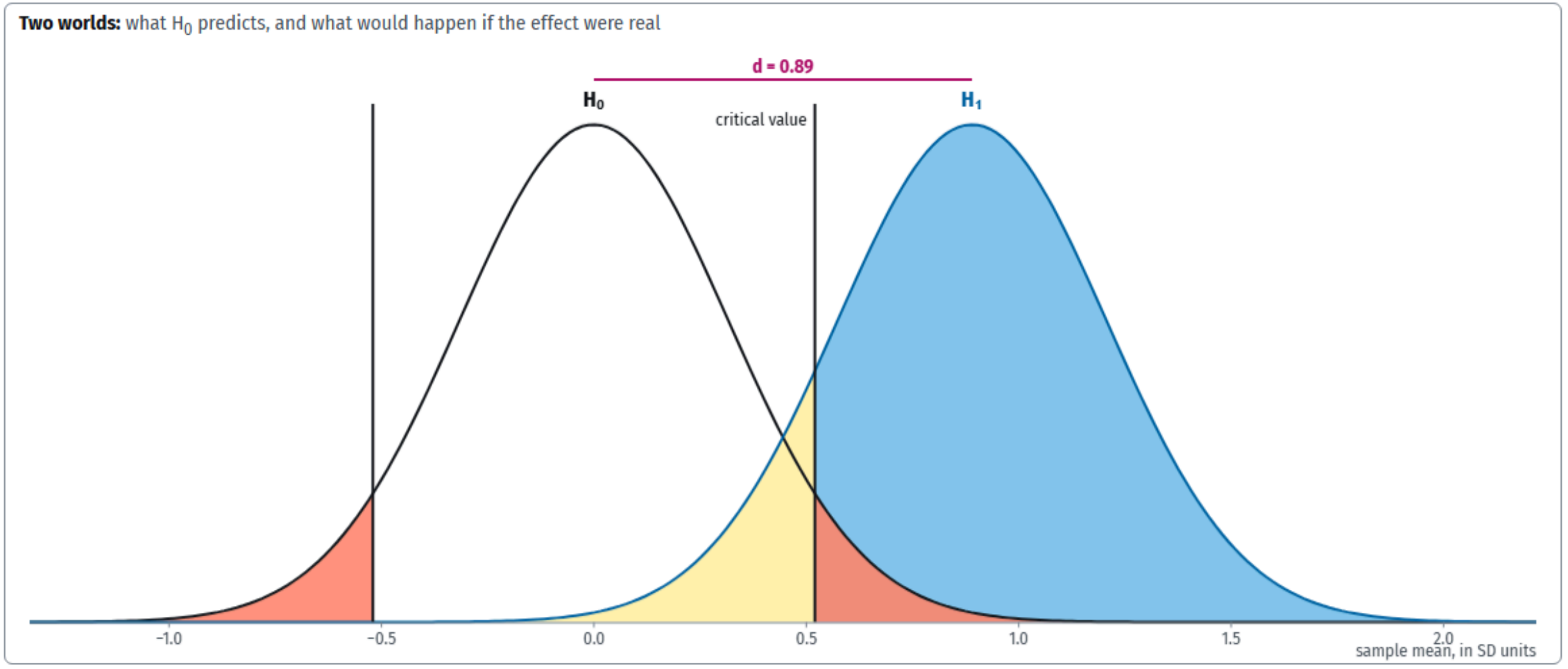
2. Use one tail: a directional test

3. Raise n: more people per sample

4. Find a larger effect: a larger d

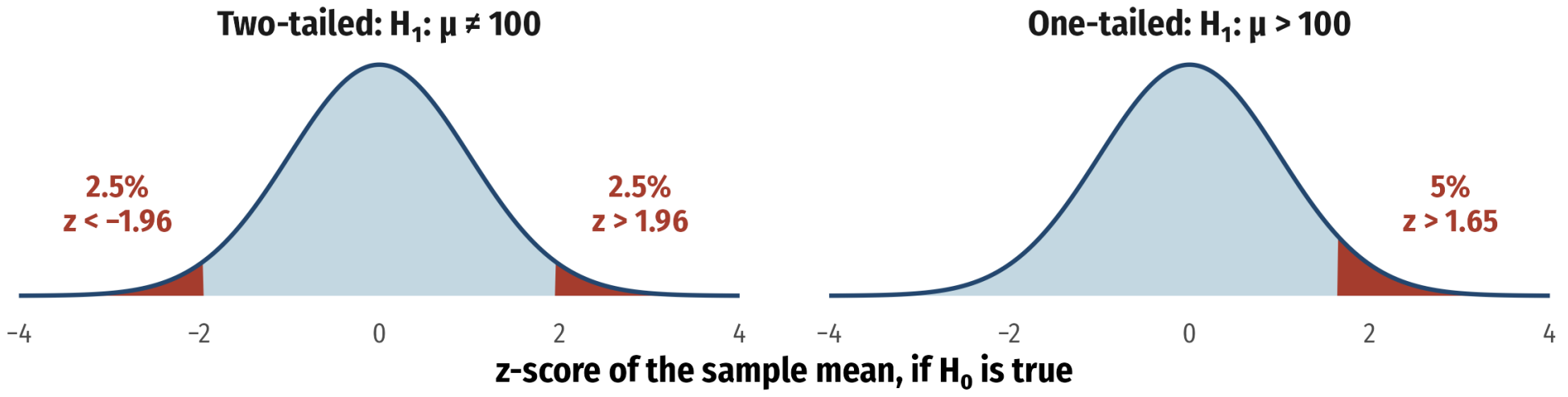
We start from d = 0.89, n = 10 and α = 5%, two-tailed, with power of 80%

Way 1: Looser α lifts power to 88%



Simulator after Kristoffer Magnusson, rpsychologist.com.

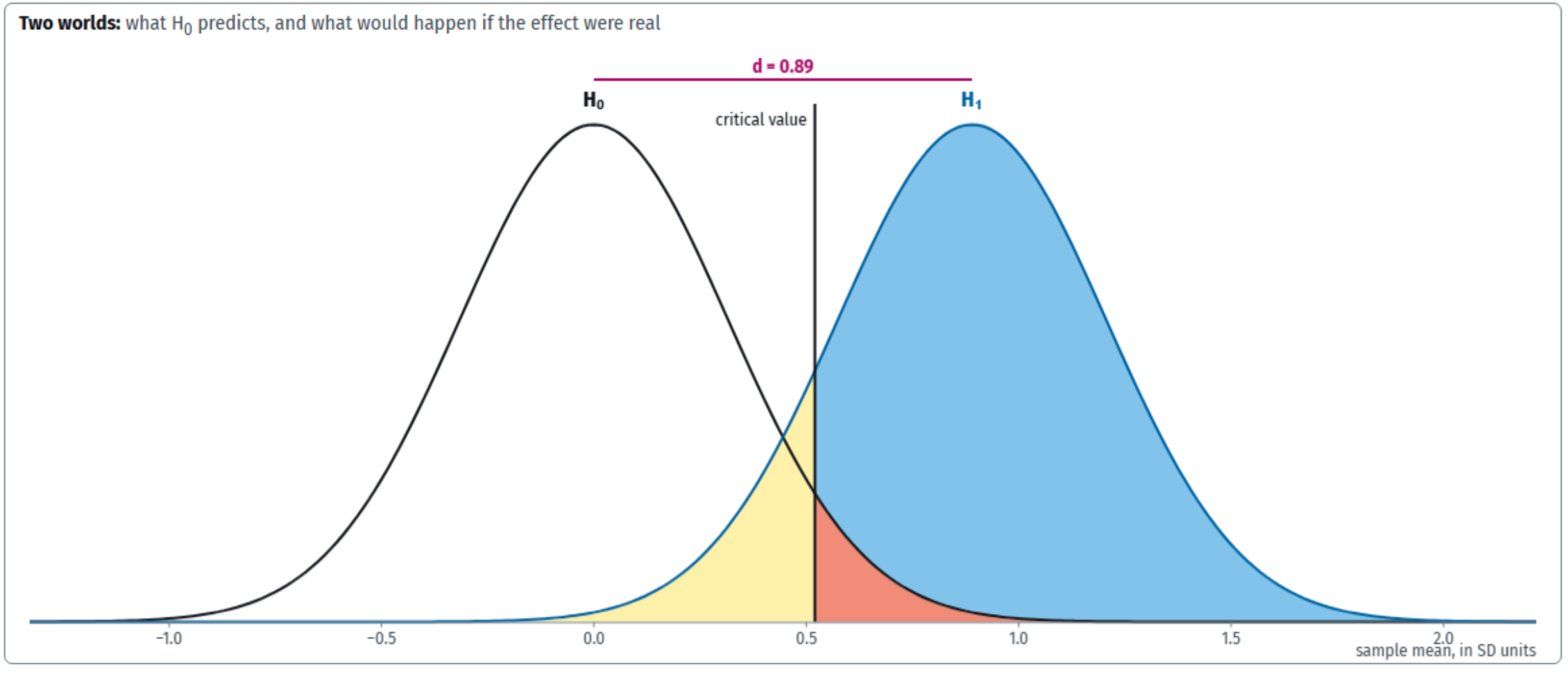
A directional test puts all of α in one tail



H_0: μ ≤ 100 and H_1: μ > 100. α stays at 5%, so the cutoff moves in

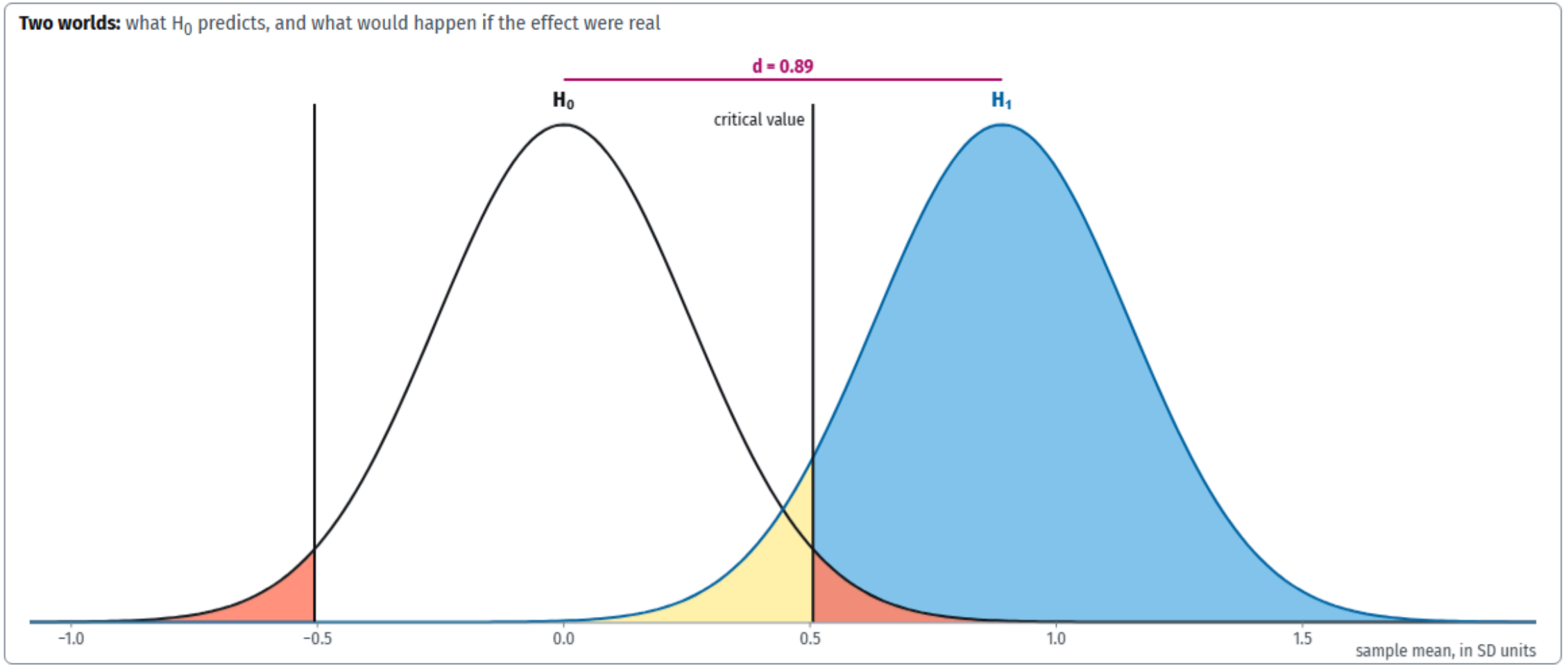
Choose the direction before seeing the data. An effect the other way can't be found

Way 2: One tail lifts power to 88%



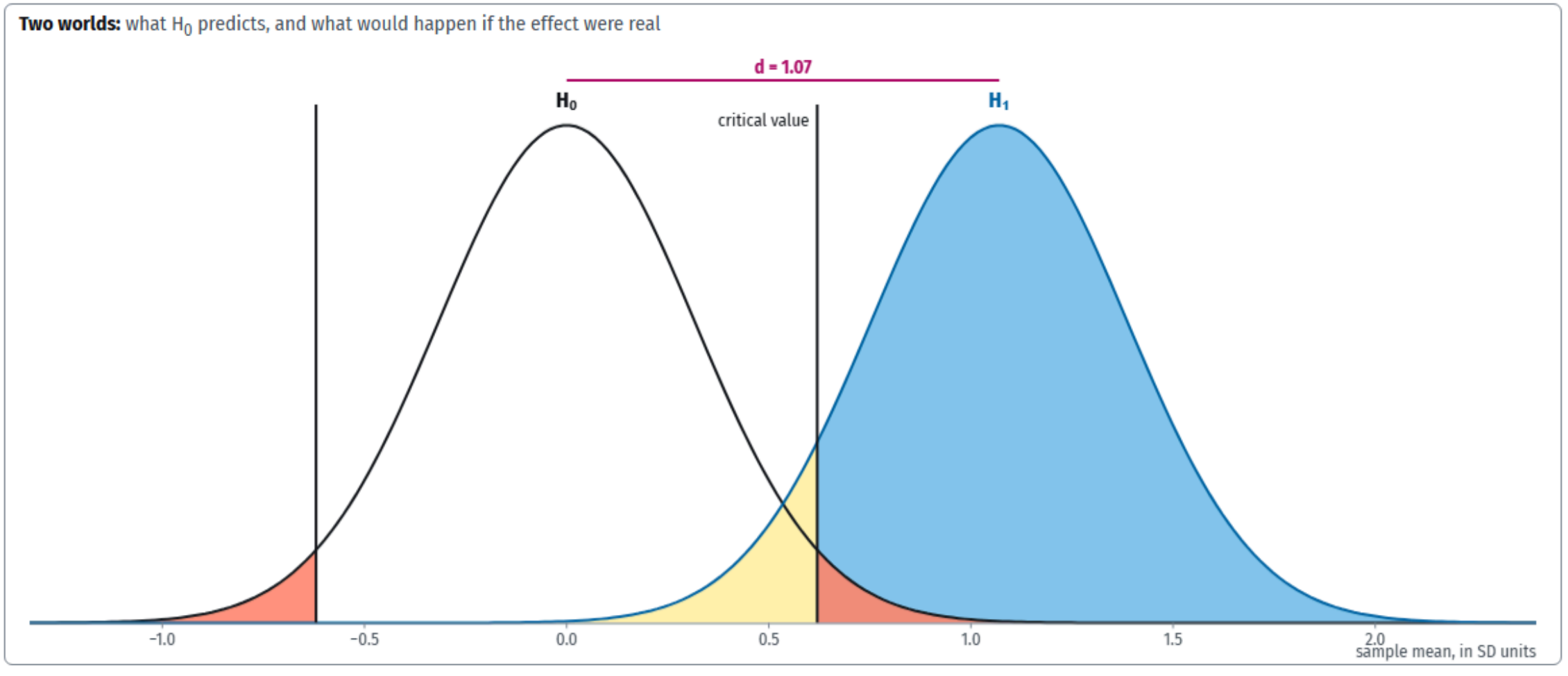
Simulator after Kristoffer Magnusson, rpsychologist.com.

Way 3: Larger n lifts power to 93%



Simulator after Kristoffer Magnusson, rpsychologist.com.

Way 4: Larger d lifts power to 92%



Simulator after Kristoffer Magnusson, rpsychologist.com.

The four ways, side by side

Change	Power	What it costs
Start: d = 0.89, n = 10, α = 5%, two-tailed	80%	
1. Loosen α to 10%	88%	Twice the Type 1 errors
2. One-tailed test	88%	Blind to an effect in the other direction
3. n = 15	93%	More students, more money
4. d = 1.07	92%	Needs a stronger program or a better-matched test

Only ways 3 and 4 raise power without touching what we count as evidence

Predict it: Rescue a weak study

A real but small effect: d = 0.20. With n = 10 and α = 5%, power is only 10%

Which rescue helps more?

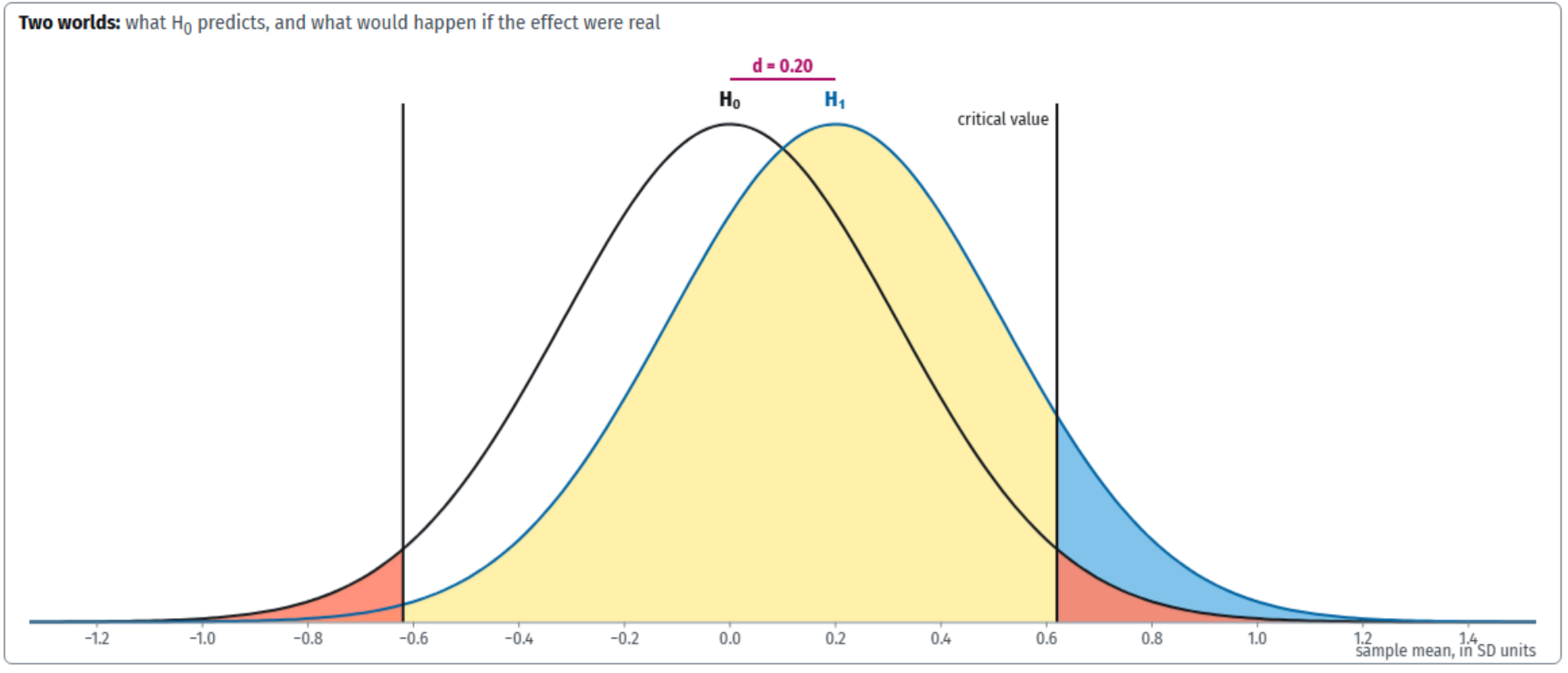
A Loosen α to 10%

B Raise n from 10 to 40

C They help about the same

Then the Power tab decides

Predict it: Move the sliders



α = 10% gives 17%. n = 40 gives 24%. Both are far from 80%

Simulator after Kristoffer Magnusson, rpsychologist.com.

Planning a study, reading a result

Two questions before any data

The book finds the power of one test. Researchers usually fix power at 80% and turn the question around

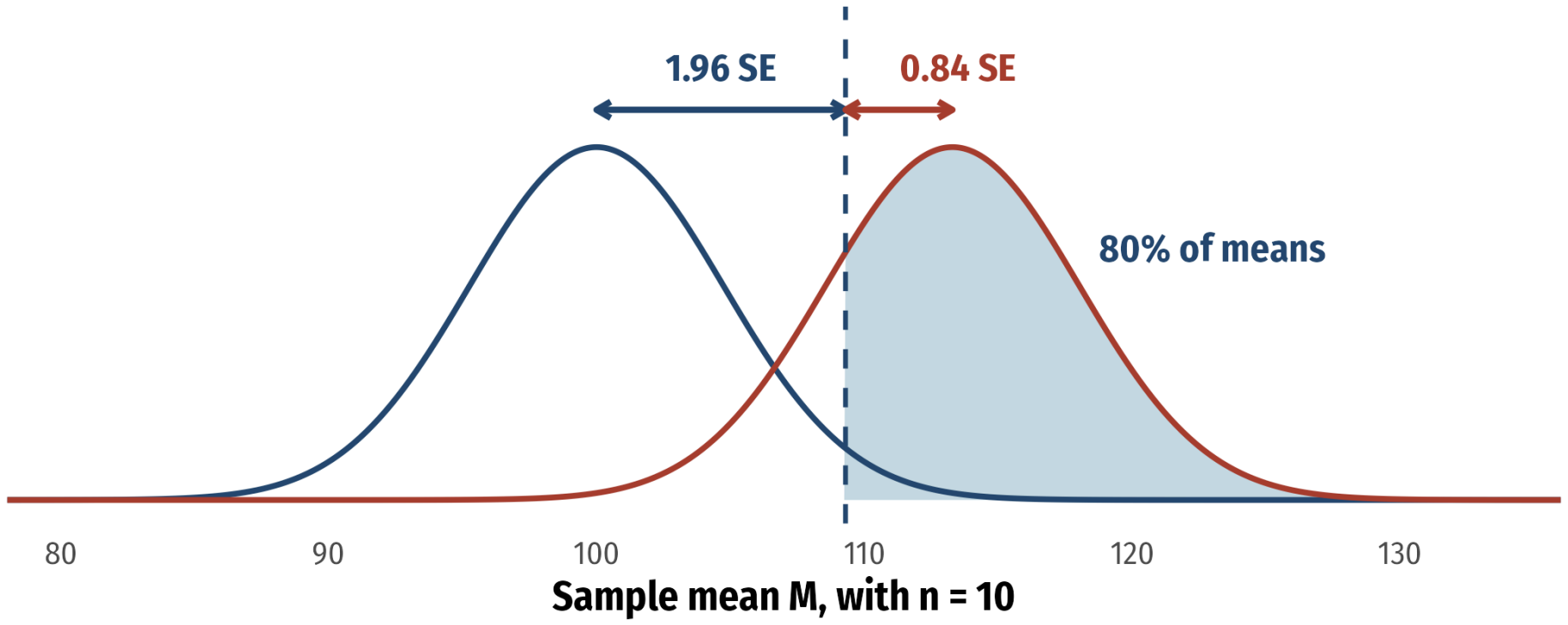
I know my n

How small an effect can I detect with 80% power? This is the **minimal detectable effect**

I know my effect

How many people do I need for 80% power? This is the **required sample size**

Where 2.80 comes from



For 80% power, μ must sit 1.96 + 0.84 = 2.80 standard errors from 100

Two formulas, one idea

Minimal detectable effect

$$d = \frac{1.96 + 0.84}{\sqrt{n}}$$

With n = 10: 2.80 / √10 = **0.89**

Required sample size

$$n = \left(\frac{1.96 + 0.84}{d}\right)^2$$

With d = 0.89: (2.80 / 0.89)² = **9.9**, so 10

That's why n = 10 had 80% power for d = 0.89. These use d in SD units, so σ = 1

In education, an effect of 0.5 SD is rare

Which studies	Mean	Median	90th percentile
All	0.16	0.10	0.47
Math tests	0.11	0.07	0.37
100 students or fewer	0.30	0.24	0.77
More than 2,000 students	0.05	0.03	0.17

Effects from randomized trials of education programs, in SD units. Bigger studies find smaller effects

Mindspark's endline gap, 2.69 points, is about 0.46 SD: very large here

Please ignore the book's Table 8.2, where 0.2 is "small" and 0.8 is "large"

Effects: Kraft (2020), *Educational Researcher*, 49(4). Mindspark: Muralidharan, Singh, and Ganimian (2019).

Statistical significance isn't importance

Sample size	Effect	z	p	Decision
100	d = 0.05	0.50	0.617	Maintain H_0
1,000	d = 0.05	1.58	0.114	Maintain H_0
10,000	d = 0.05	5.00	< 0.001	Reject H_0

Same tiny effect, opposite verdicts. With a large n, the sampling distribution gets so narrow that even small effects are significant

Significance says an effect is detectable. Its size says if it matters

Try it: Plan two studies

Use power of 80% and α = 0.05, two-tailed. The multiplier is 2.80

Study A

You can test 25 students. What's the smallest effect you can detect? Is it realistic, by Kraft's table?

Study B

You expect Kraft's median effect, d = 0.10. How many students do you need?

Work alone for two minutes. Compare with a neighbor. Then we vote

Effects: Kraft (2020), *Educational Researcher*, 49(4).

Try it: Answer

Study A

$2.80 \;/\; \sqrt{25} =$ **0.56** SD

That's above the 90th percentile of all studies in Kraft's table. This study would miss most real programs

Study B

$(2.80 \;/\; 0.10)^2 =$ **784** students

Realistic effects need big samples

Small studies mostly find nothing, and when they do find something, it tends to be overstated

Effects: Kraft (2020), *Educational Researcher*, 49(4).

What you can do now

Run the four steps: state H_0, set criteria, observe a sample, reject or maintain

Say what a p-value is, and what it isn't

Tell a Type 1 error from a Type 2 error, and locate α, β and power on two curves

Compute Cohen's d, and name four ways to raise power

Find a minimal detectable effect and a required sample size

Explain why statistical significance isn't importance

Next steps

Before Thursday

Read sections 9-1, 9-2 and 9-4 on the t-test, for when σ is unknown

This Friday: Lab 5

Power and the one-sample t-test. Bring a calculator and a laptop

Mastery and Problem Set for Chapter 8 are due Sunday, February 7 at 6 PM. The midterm is Tuesday, February 16