J-PAL
ABDUL LATIF JAMEEL POVERTY ACTION LAB

Mechanics of Power Calculations

Course Overview

1. Why Evaluate
2. Theory of Change & Measurement
3. Why & When to Randomize
4. How to Randomize
5. Sample Size & Power
    1. Essentials of Power
    2. Mechanics of Power (you are here!)
6. Randomized Evaluation from Start to Finish
7. Threats & Analysis
8. Ethical Considerations
9. Generalizing & Applying Evidence

Power tracks

- **Essentials of Sample Size and Power**: The lecture will cover the intuition behind power calculations and go over some basic principles for determining a study size that minimizes the probability of false negatives. It is aimed at policymakers and practitioners who wish to understand the essentials of power and how various components can be tweaked when designing a study.

- **Mechanics of Power Calculations**: The lecture is designed for participants who are looking to discuss statistical power in more depth and may be planning on conducting power calculations in the near future. The lecture provides the statistical framework for power, introduces its components, and provides practical guidance for power and sample size calculations. The lecture also includes a short exercise. This lecture might be right for you if you:
  - Have taken at least one class on probability theory, statistics, or econometrics
  - Have at least some experience working with data
  - Have at least some experience reading academic literature

What is statistical power?

Learning objectives

- Understand how the estimated effect size depends on the specific sample
- Understand intuitively what power is and how it relates to the probability of making false positive (type 1) and false negative (type II) errors
- Understand technically how the power of a study is derived, how it is calculated, and what components of a study's design affect its power
- Internalize the importance of doing power calculations (early)
- Feel equipped to conduct preliminary power calculations and sensitivity analyses

Outline

I. Motivation

II. Hypothesis testing and statistical power

III. Power calculations

IV. Power in cluster-randomized studies

V. Power calculations in practice

VI. Example (if time)

Outline

I. **Motivation**

II. Hypothesis testing and statistical power

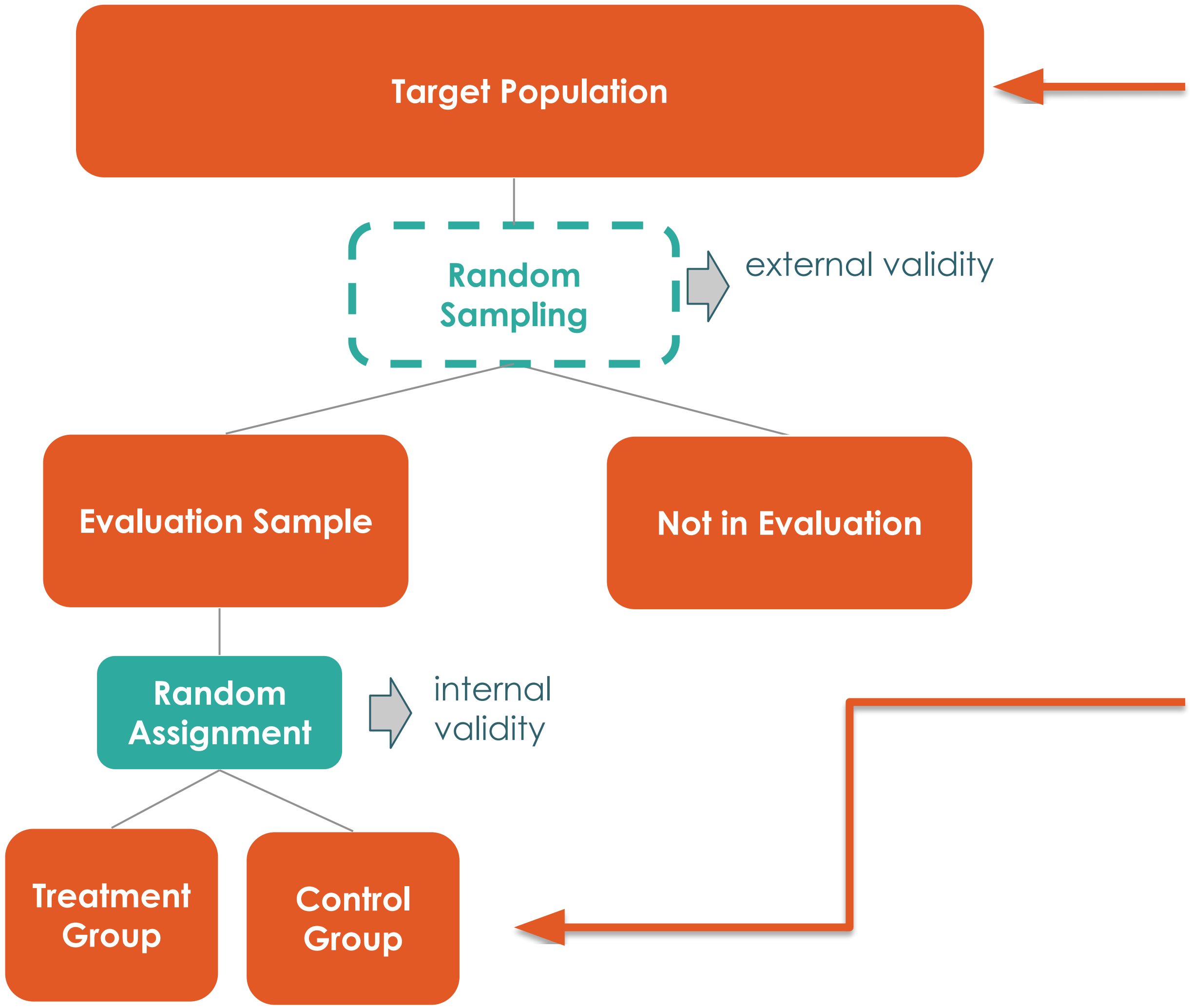
III. Power calculations

IV. Power in cluster-randomized studies

V. Power calculations in practice

VI. Example (if time)

Estimating the true treatment effect with an experiment



True treatment effect (β): the true population difference in the outcome with and without the program

- Fundamentally unknowable

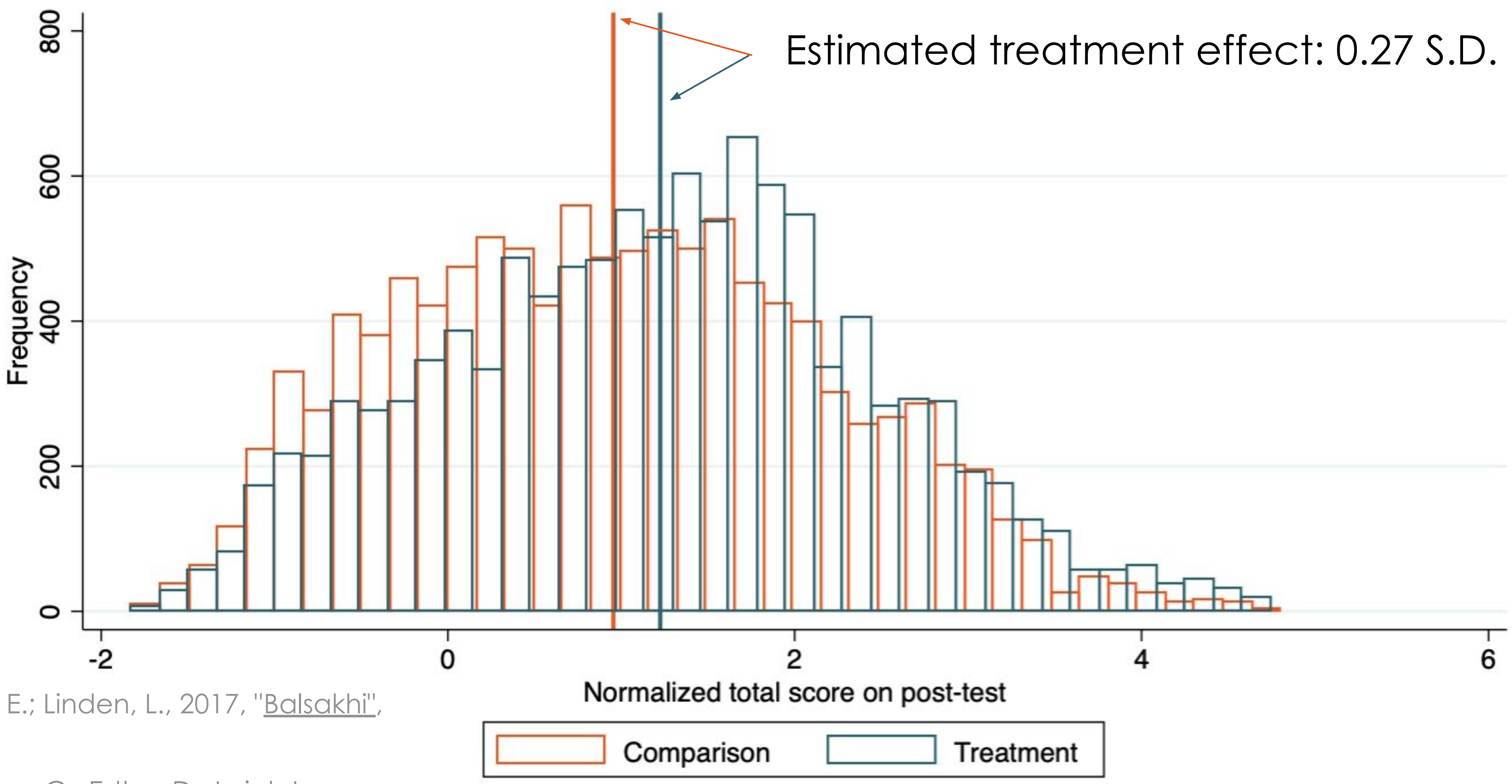
Estimated treatment effect ($\hat{\beta}$): the sample difference in the outcome between the treatment and comparison group

- The estimated effect depends on the specific sample in your RCT
- The estimated effect depends less on the sample the larger the sample size

Empirical example: Balsakhi tutoring program

Study: Balsakhi remedial tutoring program in India

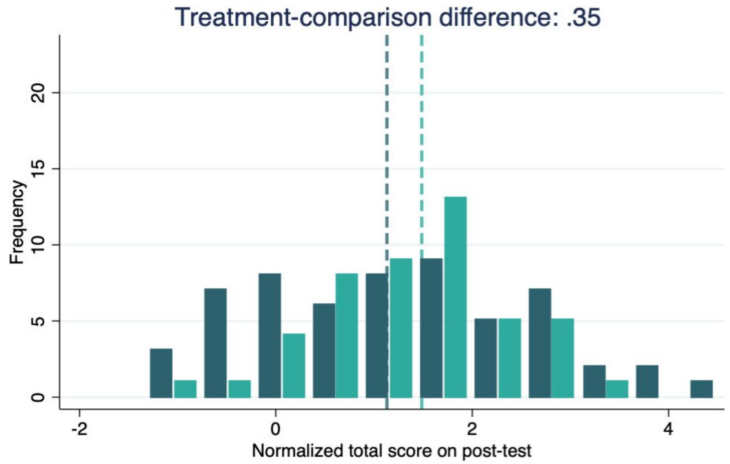
Sample size: More than 23,000 students

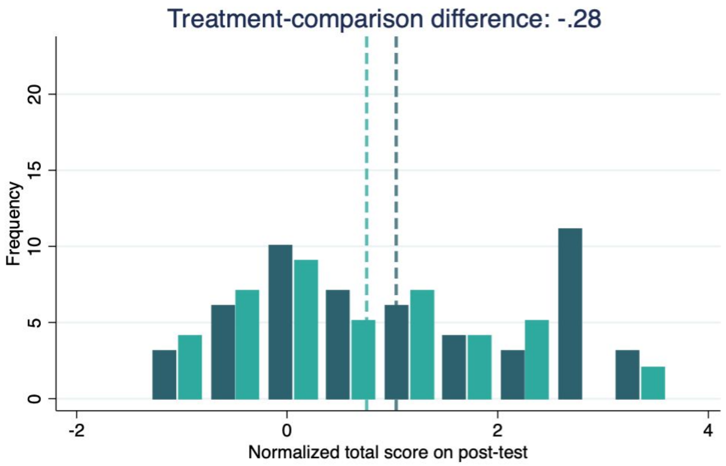


Source: Banerjee, A., Cole, S.; Duflo, E.; Linden, L., 2017, "Balsakhi", Harvard Dataverse

Research publication: Abhijit B., Shawn C., Esther D., Leigh L.; "Remedying Education: Evidence from Two Randomized Experiments in India", The Quarterly Journal of Economics 122(3), 1235–1264.

Different random samples from the same population lead to different treatment effect size estimates

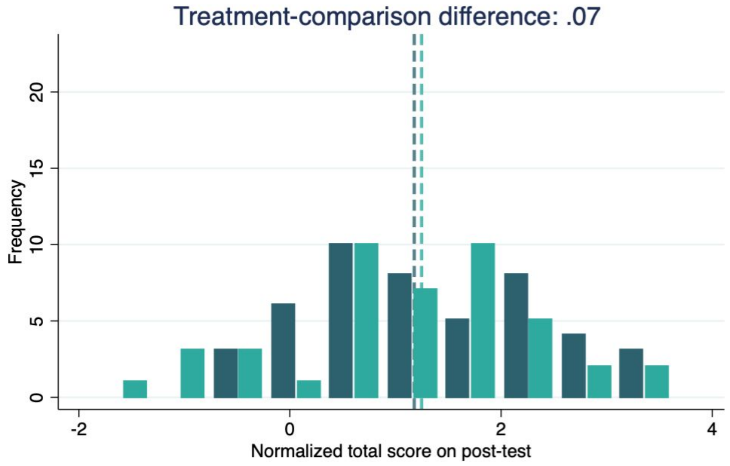




■ Comparison

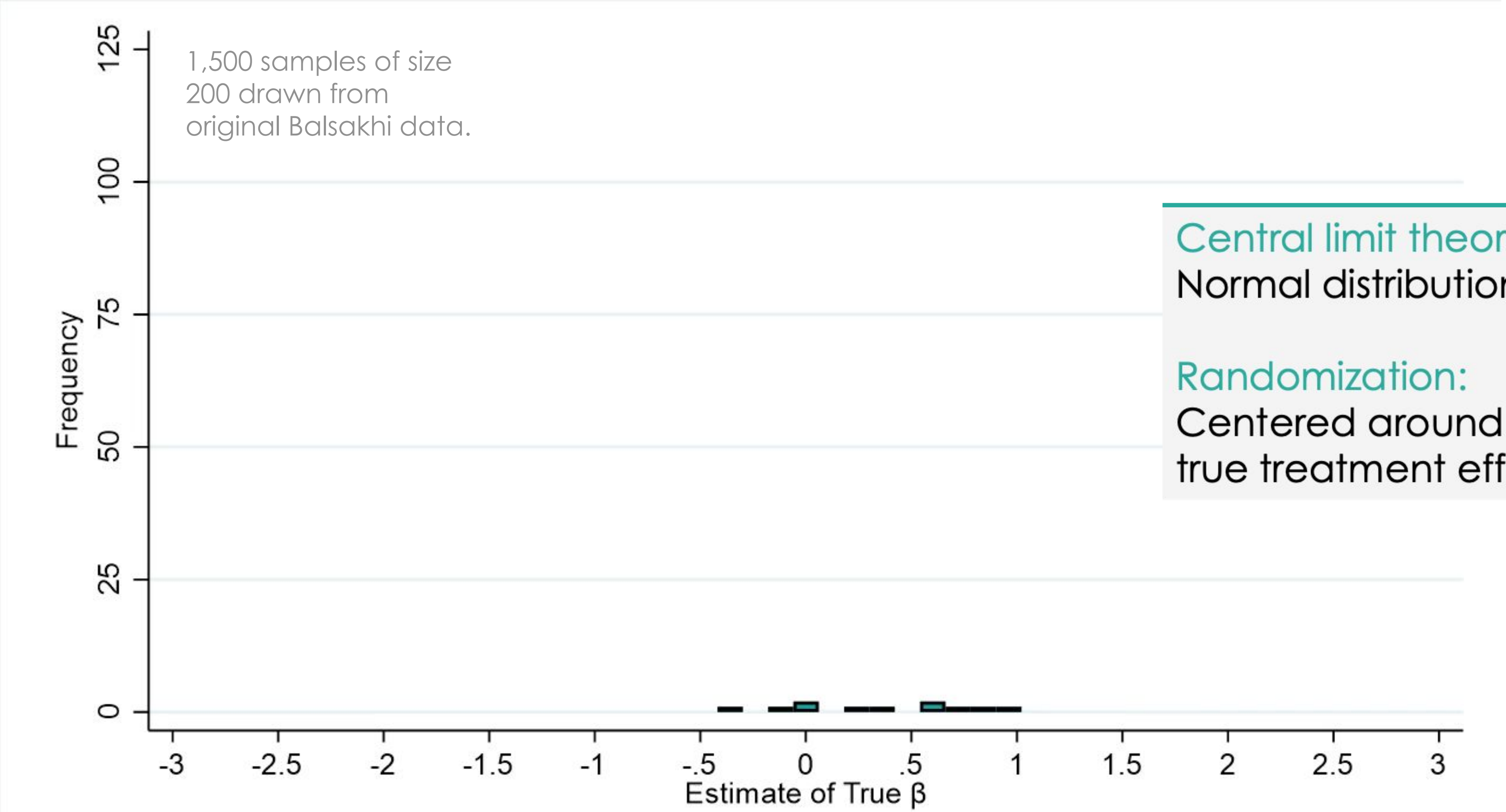
■ Treatment

Samples of size 200 drawn from original Balsakhi data.



Challenge: Is the difference between groups due to chance variation or an effect of the program?

Many samples: a *sampling distribution* of estimates

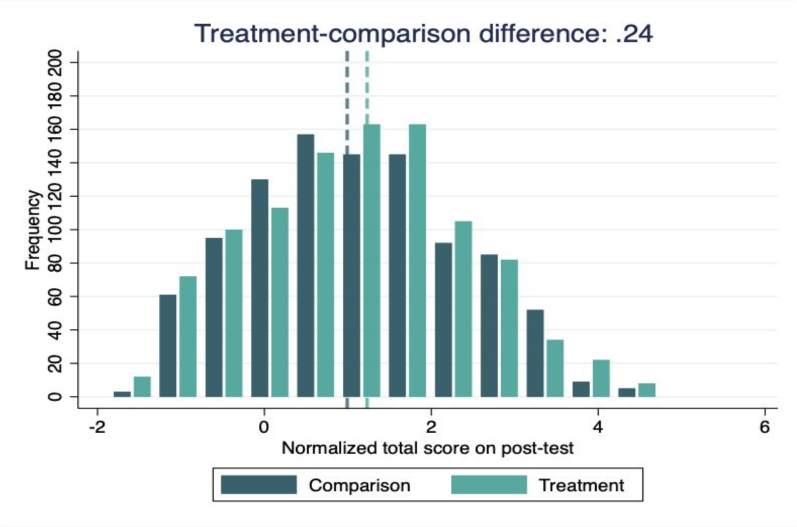


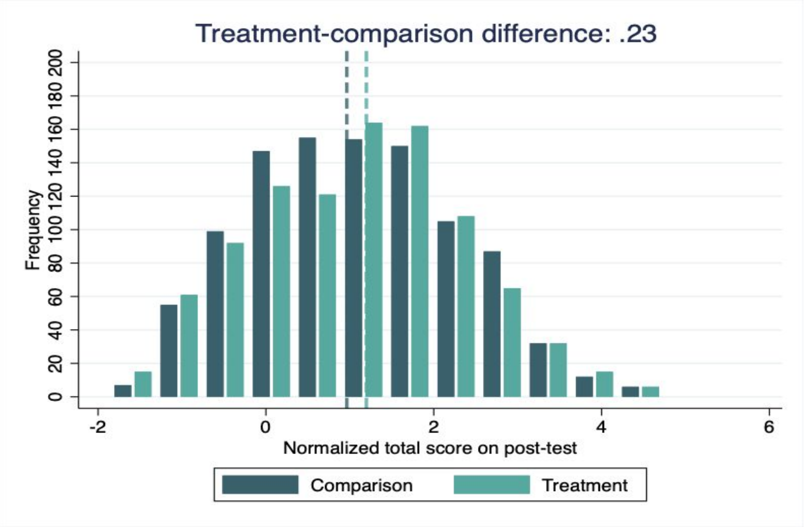
Central limit theorem:
Normal distribution

Randomization:
Centered around the true treatment effect, β

What do you think happens if we increase the sample size?

Larger samples lead to less random variation in treatment effects

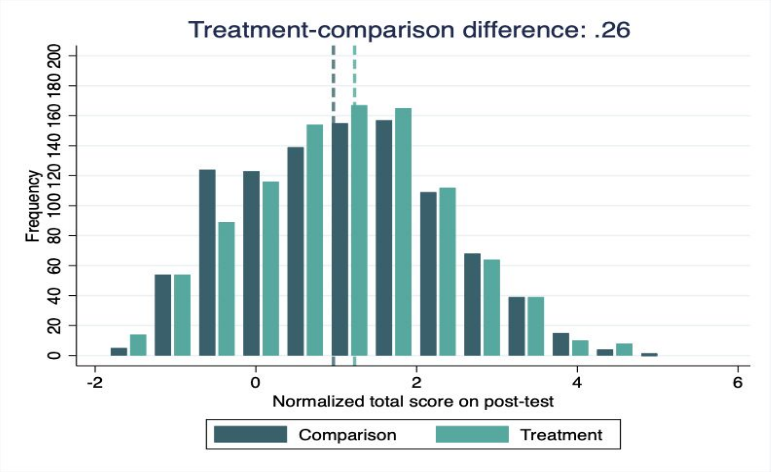

Treatment-comparison difference: .24


Treatment-comparison difference: .23

🟦 Comparison

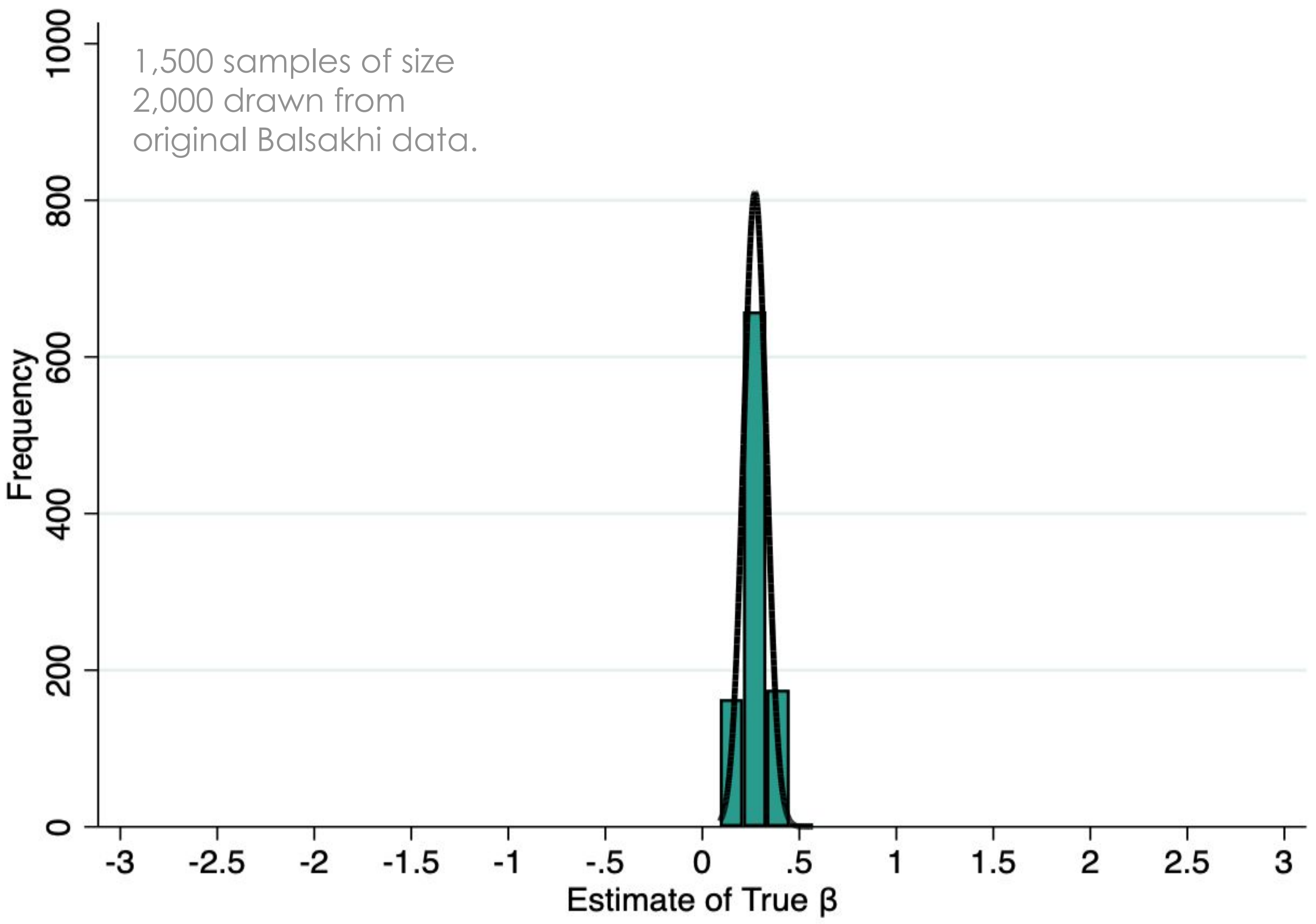
🟩 Treatment

Samples of size 2,000 drawn from original Balsakhi data.

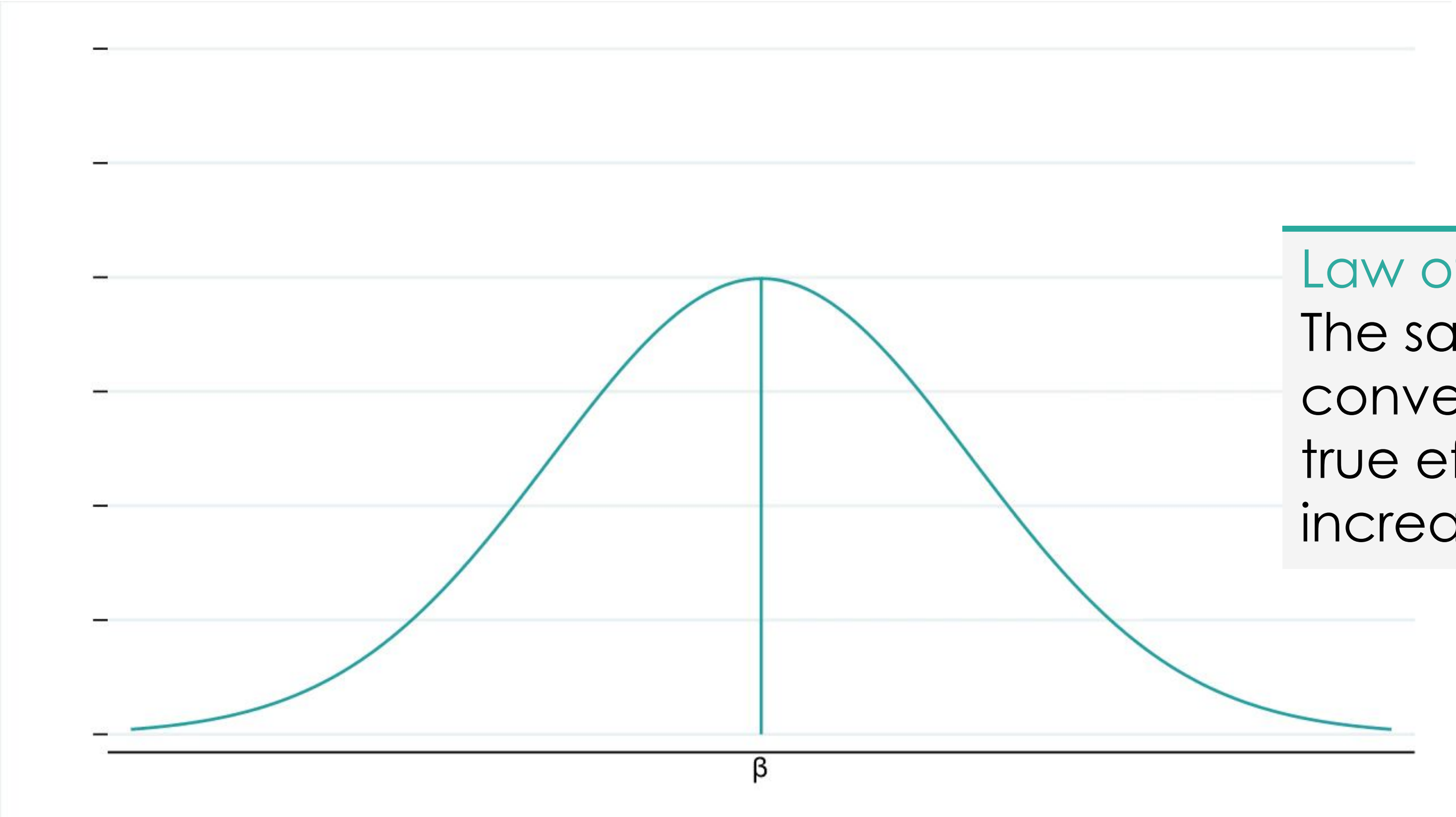

Treatment-comparison difference: .26

When the sample size is larger, any observed difference is more likely to be caused by the program than sampling variation.

The larger the sample, the more likely the estimated treatment effect is close to the true treatment effect



General rule: The larger the sample size, the narrower the sampling distribution…



Law of large numbers: The sampling distribution convergences to the true effect as *N* increases

… and the more likely the estimated treatment effect is close to the true treatment effect

Motivation/preview: Sample size and power

- The larger the sample, the more likely it is that the estimated treatment effect, $\hat{\beta}$, is close to the true treatment effect, β

- Goal of **power calculations**: Want to ensure the **sample size** is large enough to distinguish whether observed differences between treatment and comparison groups are due to random chance or due to a true impact of the program
  - Too small: risk overlooking a true effect
  - Too large: unnecessary use of resources

- The **power** of a study tells us something about the relationship between the sample size and the risk of overlooking true effects

Outline

I. Motivation

II. Hypothesis testing and statistical power

III. Power calculations

IV. Power in cluster-randomized studies

V. Power calculations in practice

Hypothesis testing

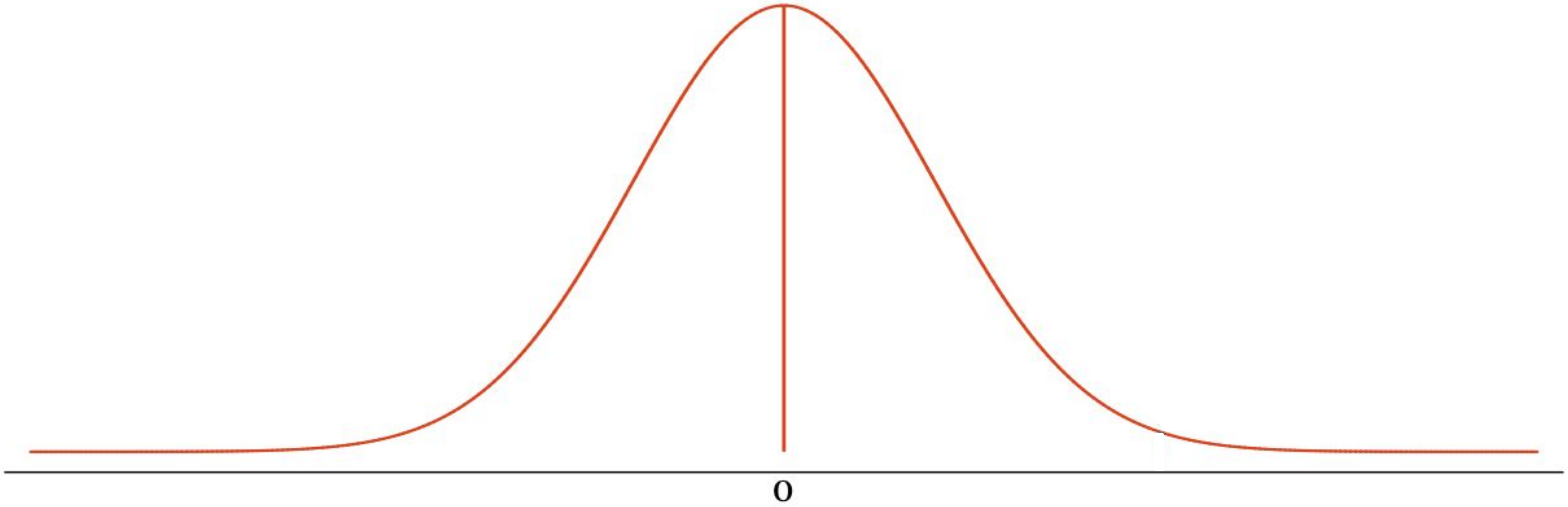
- Researchers and policymakers want to know "Is my program effective?"
  - i.e., did it change the outcome of interest?

- In hypothesis testing, you ask "what can I learn about the true treatment effect, β, by observing the estimated treatment effect, $\hat{\beta}$?"

- Hypothesis testing:
  - Start by assuming that the program did not cause any change (null hypothesis)
  - Ask: How likely is it that we would see an estimate as large as $\hat{\beta}$ in an experiment, if the true effect is actually zero?
  - If it is "very unlikely" (defined by the significance level) we reject the null hypothesis
  - If not, we fail to reject

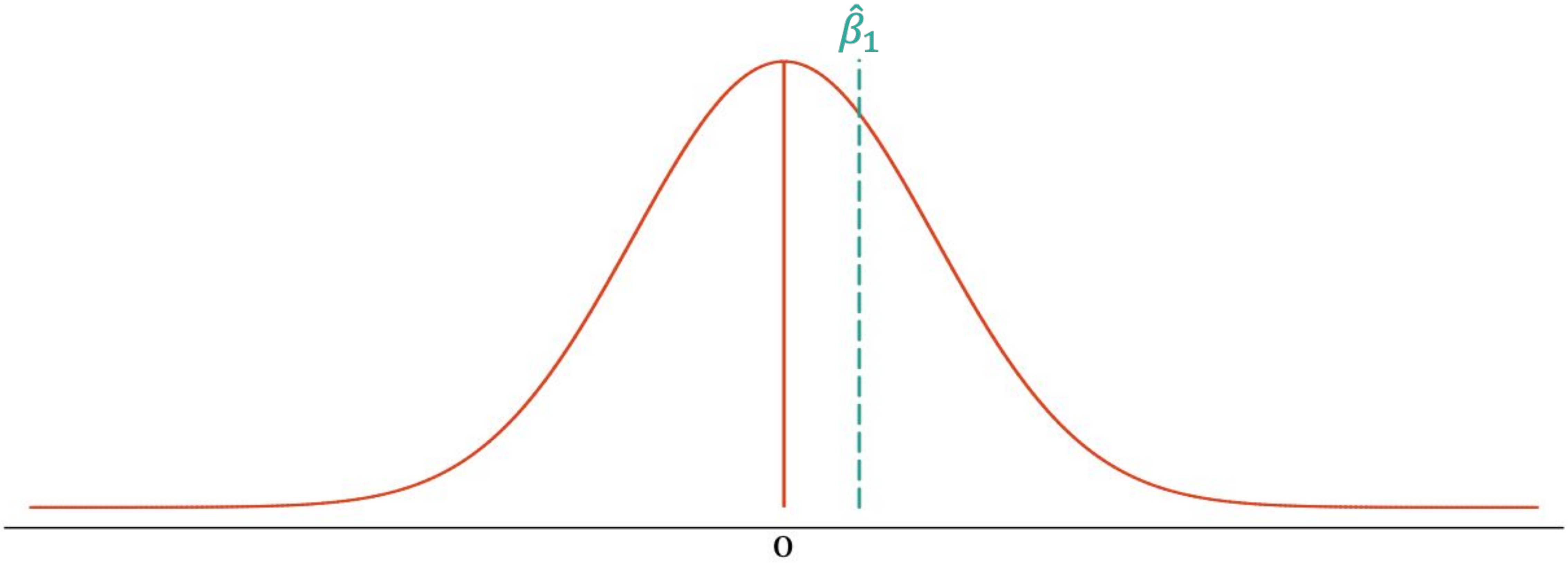
Null hypothesis:

Assume that the true treatment effect is zero



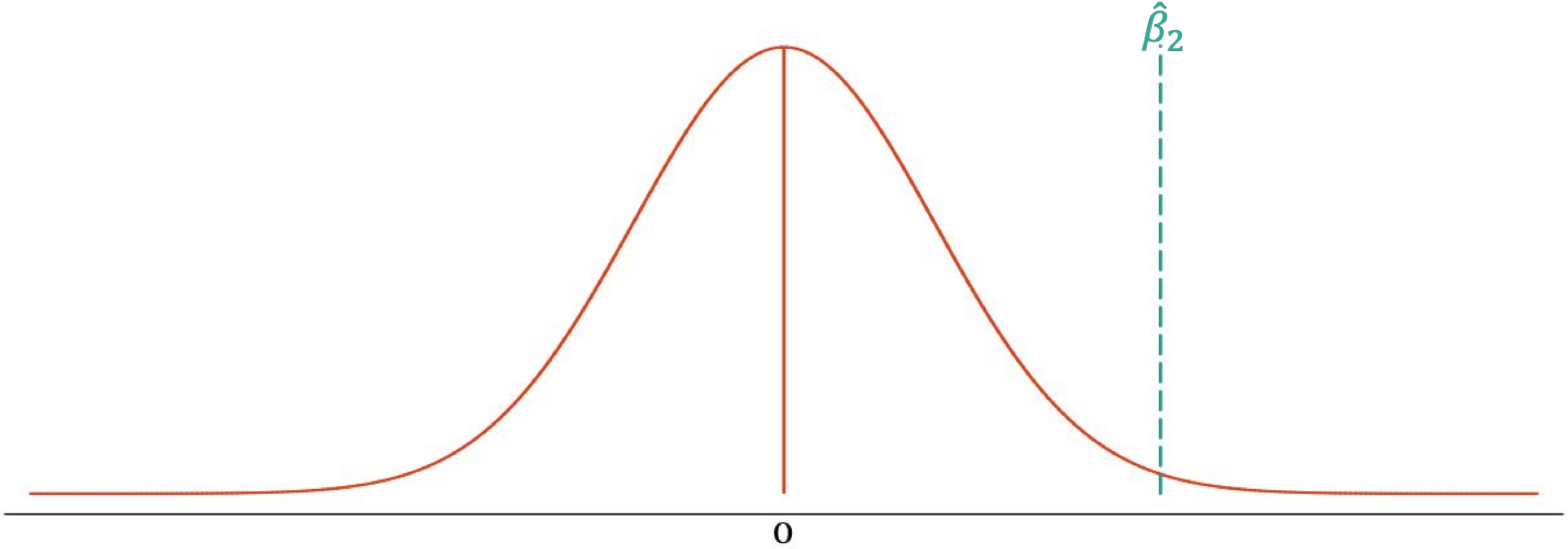
Ask "How likely is it that we would observe the treatment effect estimate, $\hat{\beta}$, if the true effect were zero?"

How likely is it to observe $\hat{\beta}_1$ under the null hypothesis?



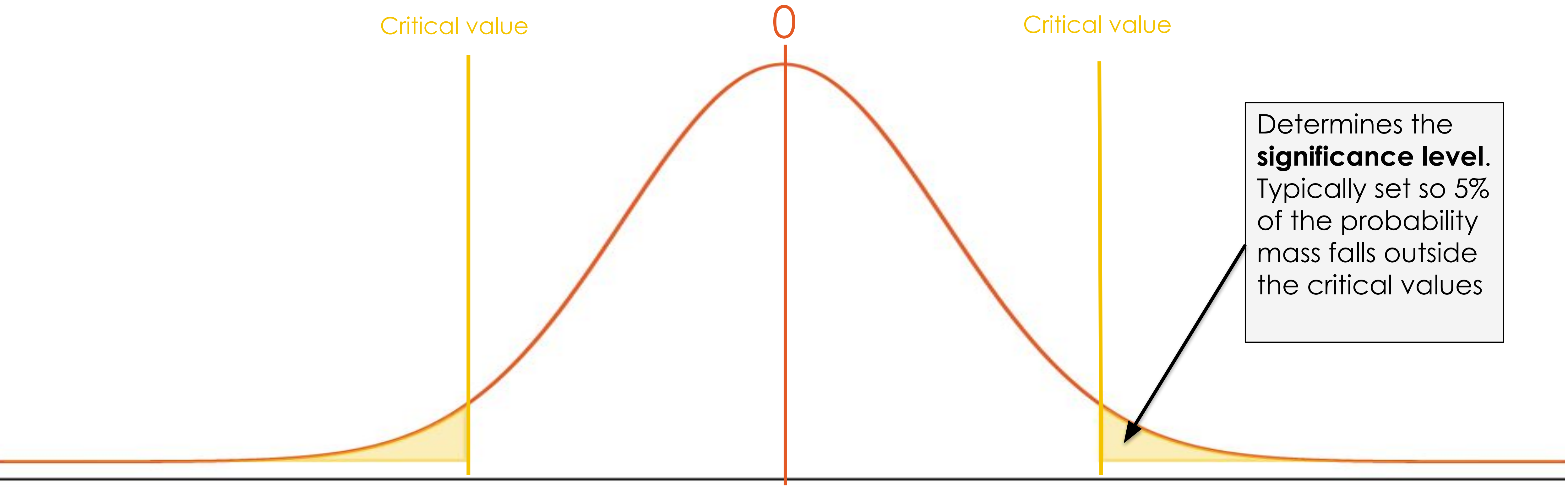
It is rather <u>likely</u> to observe estimates as large as $\hat{\beta}_1$ if the true effect were actually zero

How likely is it to observe $\hat{\beta}_2$ under the null hypothesis?



It is rather <u>unlikely</u> to observe estimates as large as $\hat{\beta}_2$ if the true effect were actually zero

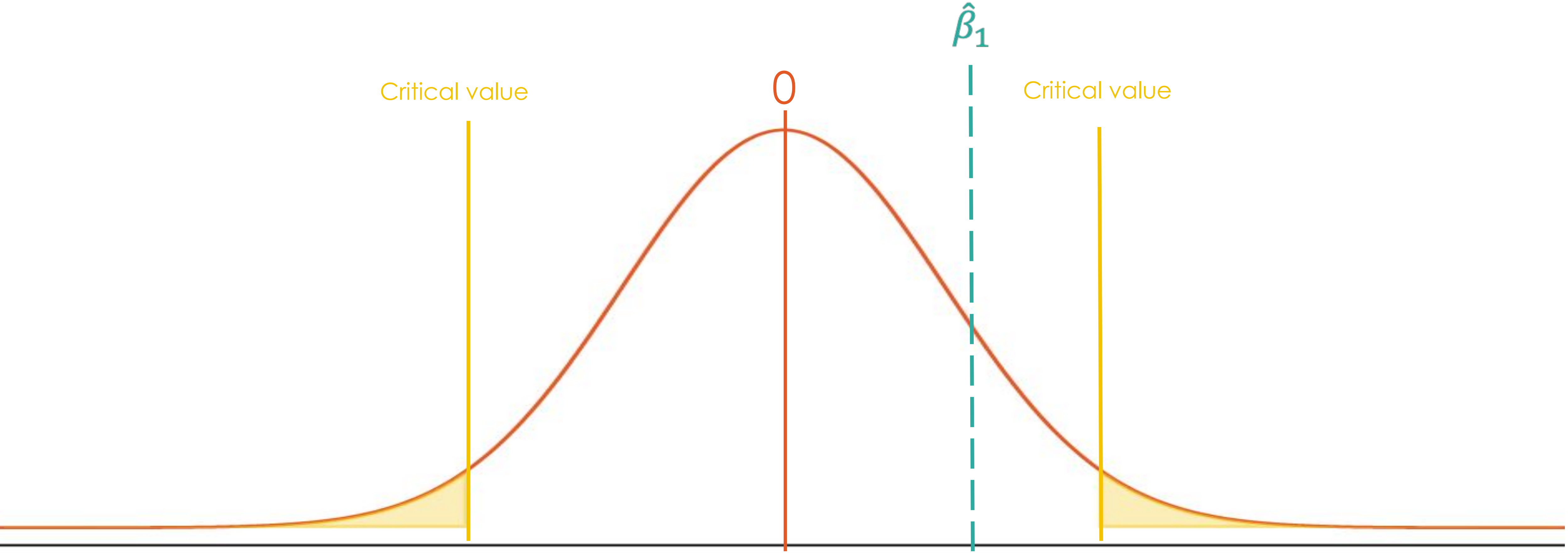
Critical values: It is “too unlikely” to observe a treatment effect outside these values if the null hypothesis is true



Determines the **significance level**. Typically set so 5% of the probability mass falls outside the critical values

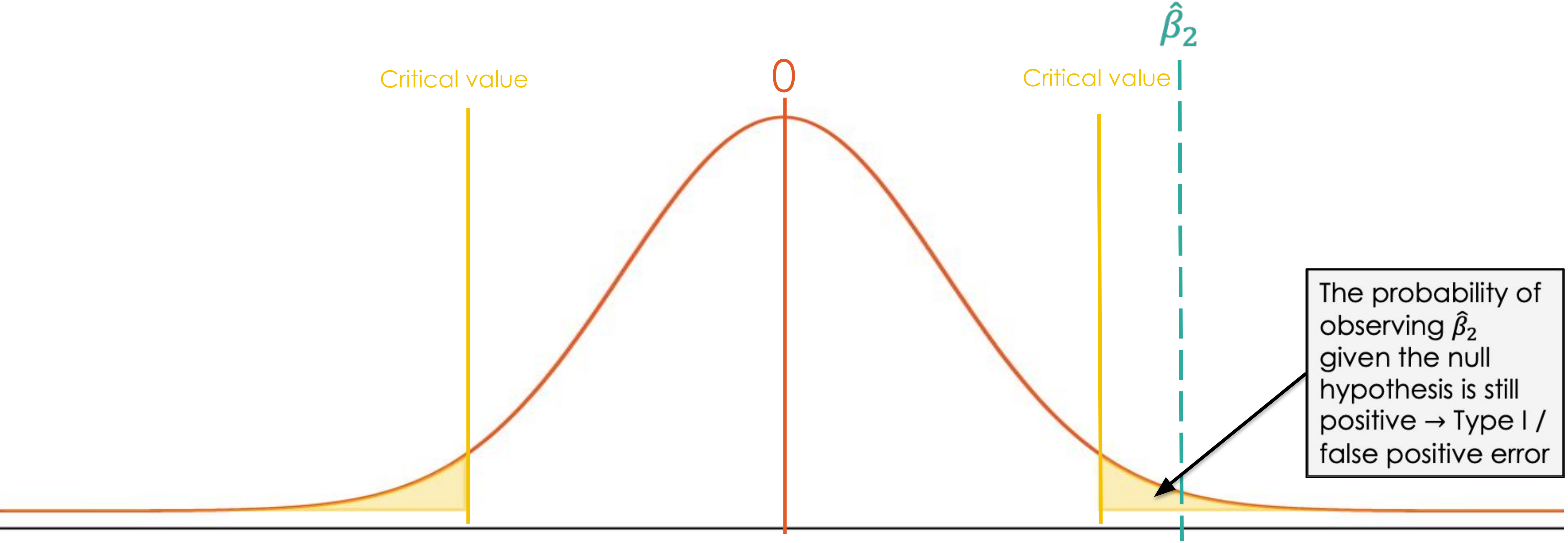
If $\hat{\beta}$ falls outside the critical values, we reject the null hypothesis

We do <u>not</u> reject the null hypothesis if we observe $\hat{\beta}_1$



"$\hat{\beta}_1$ is not statistically significantly different from zero at the 5% level"

We do reject the null hypothesis if we observe $\hat{\beta}_2$



The probability of observing $\hat{\beta}_2$ given the null hypothesis is still positive → Type I / false positive error

"$\hat{\beta}_2$ is statistically significantly different from zero at the 5% level"

Evaluation results vs. underlying reality

		True treatment effect, β	
		No impact	**Impact**
Estimated treatment effect, $\hat{\beta}$	**Conclude no impact**	True negative	
	Conclude impact	False positive/ Type 1 error (typically 5%)	True positive

Type I error (false positive)

The probability of falsely concluding that there is a treatment effect, i.e., rejecting H_0: $\beta = 0$, even if it is true. The Type I error rate is determined by the significance level.

What are some consequences of making *false positive* (Type I) errors in impact evaluations?

Is there a cost to not being willing to make *false positive* (Type I) errors in impact evaluations?

Evaluation results vs. underlying reality

True treatment effect, β

Estimated treatment effect, $\hat{\beta}$

	No impact	Impact
Conclude no impact	True negative	False negative/ Type 2 error
Conclude impact	False positive/ Type 1 error (typically 5%)	True positive

Type II error (false negative)

The probability of falsely concluding that there is no treatment effect, i.e., not rejecting H_0 even if it is not true.

What are some consequences of making *false negative* (Type II errors) in impact evaluations?

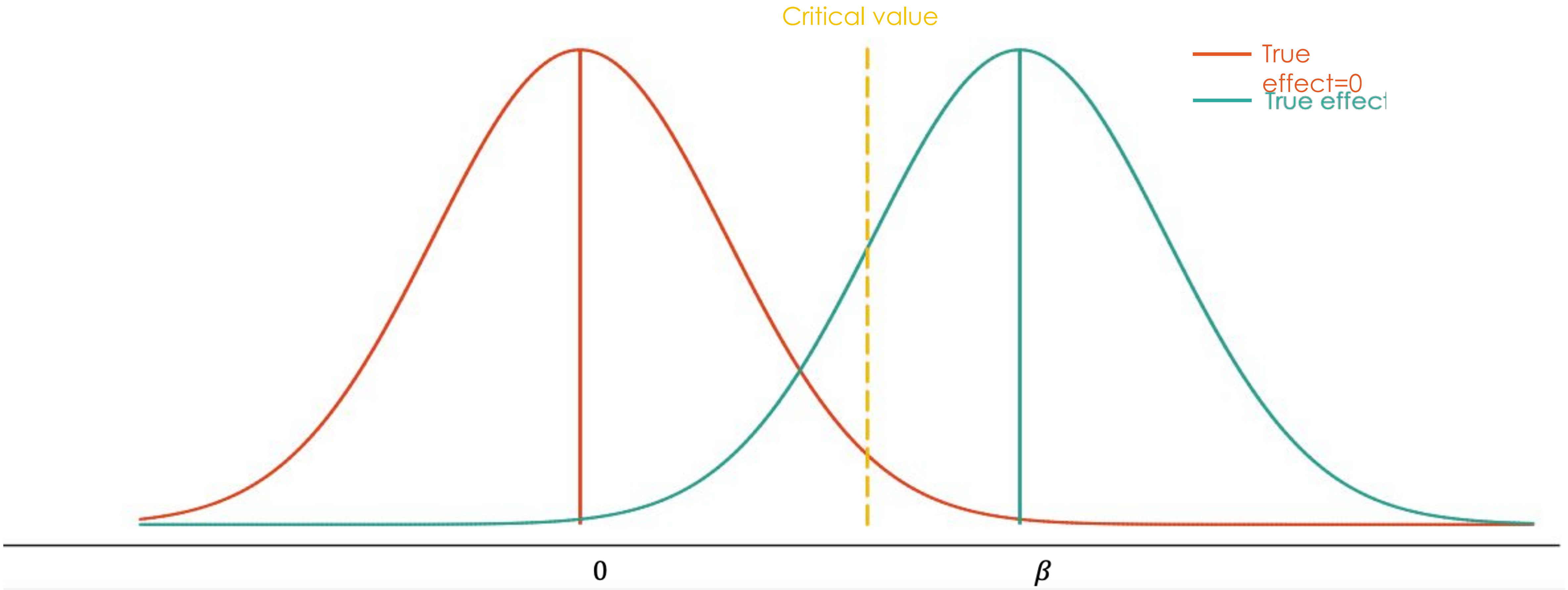
Evaluation results vs. underlying reality

		True treatment effect, β	
		No impact	**Impact**
Estimated treatment effect, $\hat{\beta}$	**Conclude no impact**	True negative	False negative/ Type 2 error
	Conclude impact	False positive/ Type 1 error (typically 5%)	True positive / Power (typically 80%)

Statistical power (true positive)

The probability of *avoiding* a Type II error, i.e., the probability of a true positive.

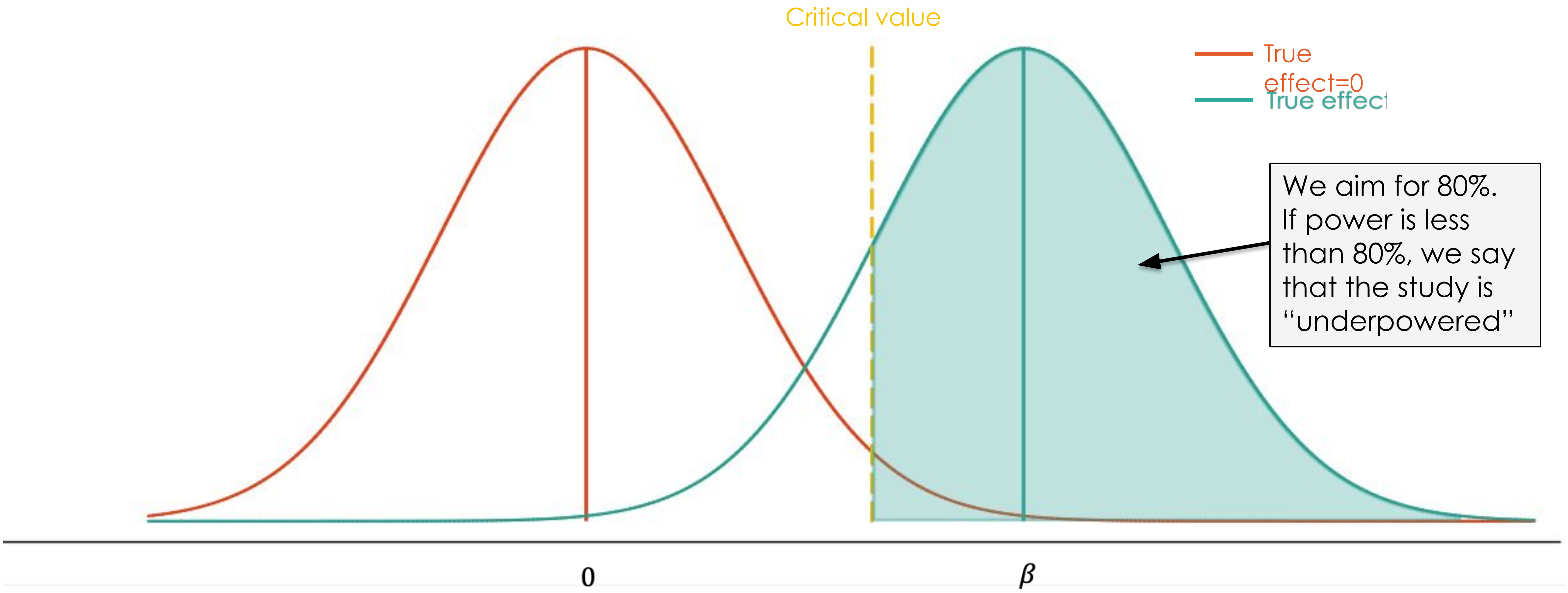
Introducing the alternative hypothesis $\beta \neq 0$



Null hypothesis: Sampling distribution centered around zero

Alternative hypothesis: Same distribution centered around $\beta \neq 0$

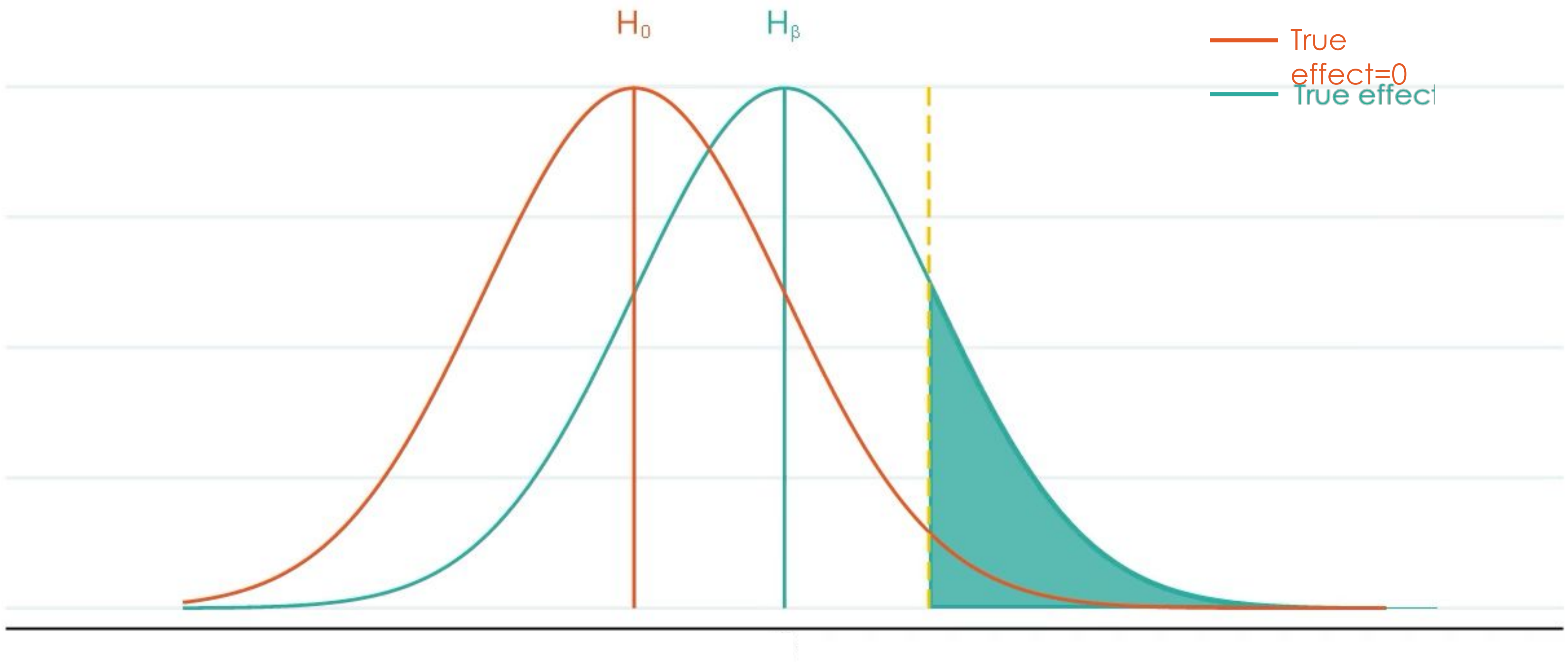
Power (true positive rate): The area to the right of the critical value under the alternative distribution



We aim for 80%. If power is less than 80%, we say that the study is "underpowered"

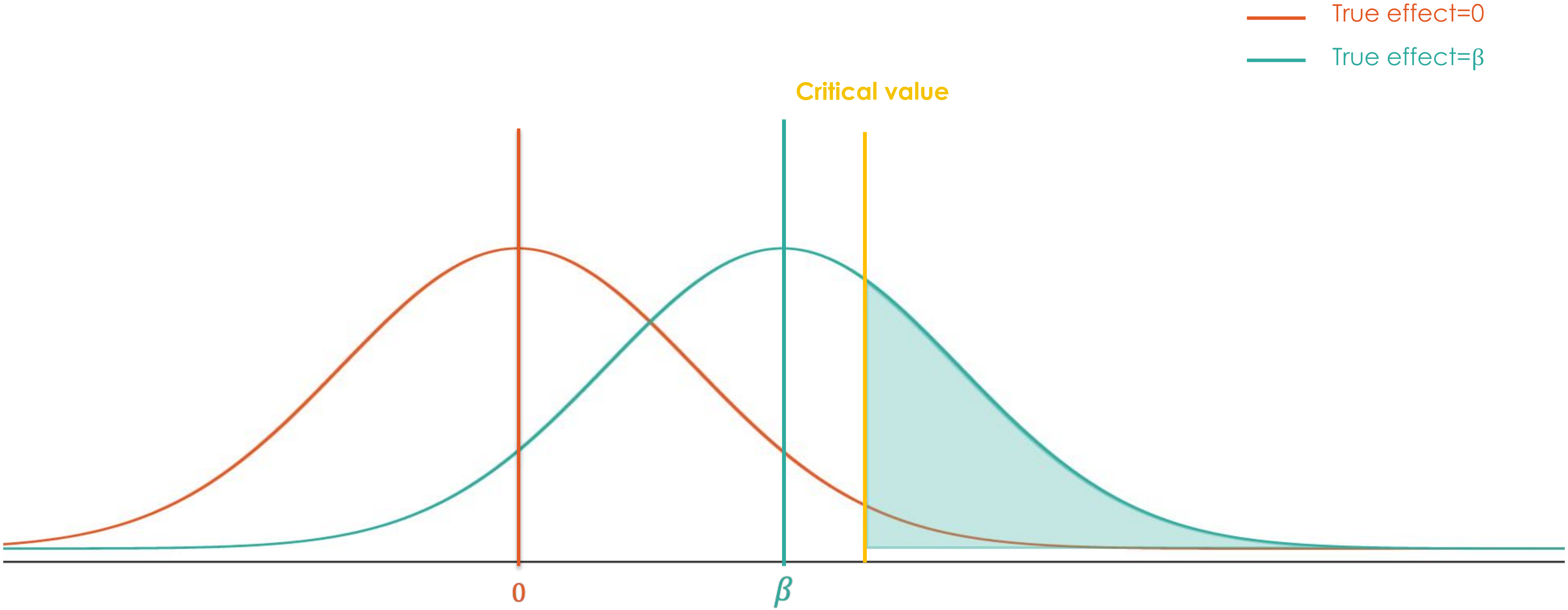
The probability of rejecting the null hypothesis when the null hypothesis is false

The larger the true effect size, the greater the power



And the easier it will be to distinguish whether a difference between the treatment and comparison groups is due to chance or an actual effect of the program.

Example of an underpowered study



Underpowered: If the probability of correctly rejecting the null hypothesis on a 5% significance level is less than 80%

What are some consequences of running under-powered studies?

Risks of running a low-powered study

- Cannot conclude whether the intervention was successful or not
- Risk of concluding that the intervention was not effective when it was
- Wasteful use of time and resources
- Will not be able to make the comparisons we want (e.g. across different treatment arms or for specific sub-groups)

Under-powered studies should be avoided

Outline

I. Motivation

II. Hypothesis testing and statistical power

III. Power calculations

IV. Power in cluster-randomized studies

V. Power calculations in practice

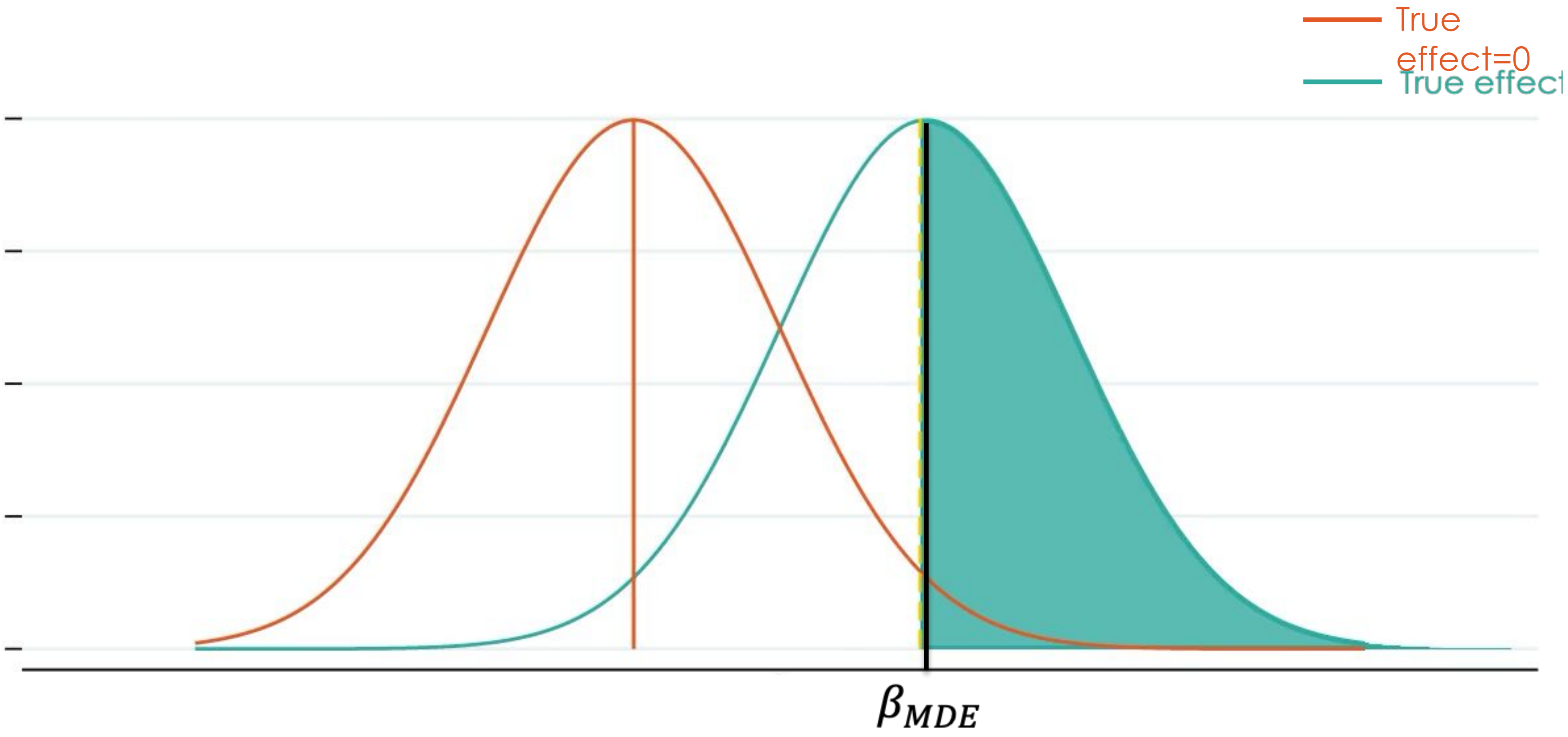
VI. Example (if time)

Power calculations: Two approaches

- **If sample size is flexible:** Calculate sample size that ensures 80% power for a given true effect size.
  - Is this sample size reasonable?
  - What sample can you reasonably recruit?
  - What sample can you reasonably manage?
  - What sample can you afford given budget constraints?

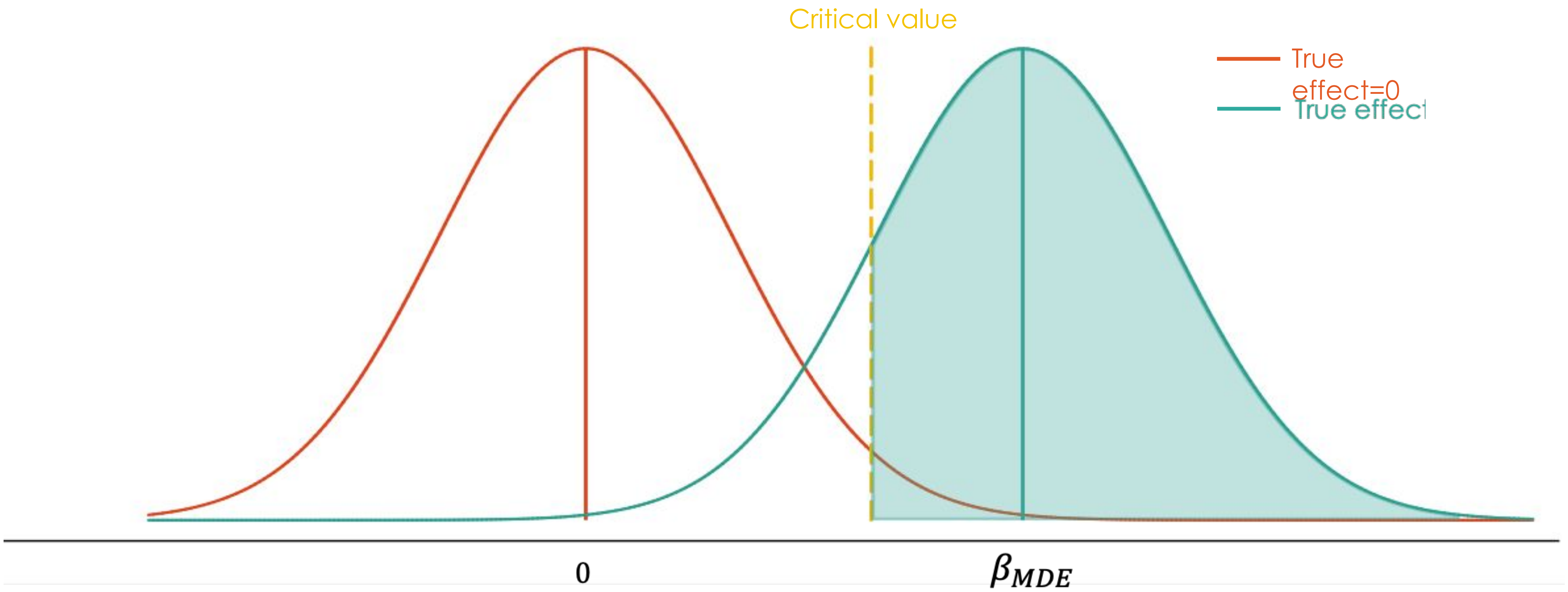
- **If sample size is fixed:** Calculate true effect size required to achieve 80% power for a given sample size.
  - Is this effect size reasonable?
  - What effects do similar studies find?
  - What effect would make the study cost-effective?
  - What effect would be required to be considered for scale-up?

Calculating required sample size for a given effect size



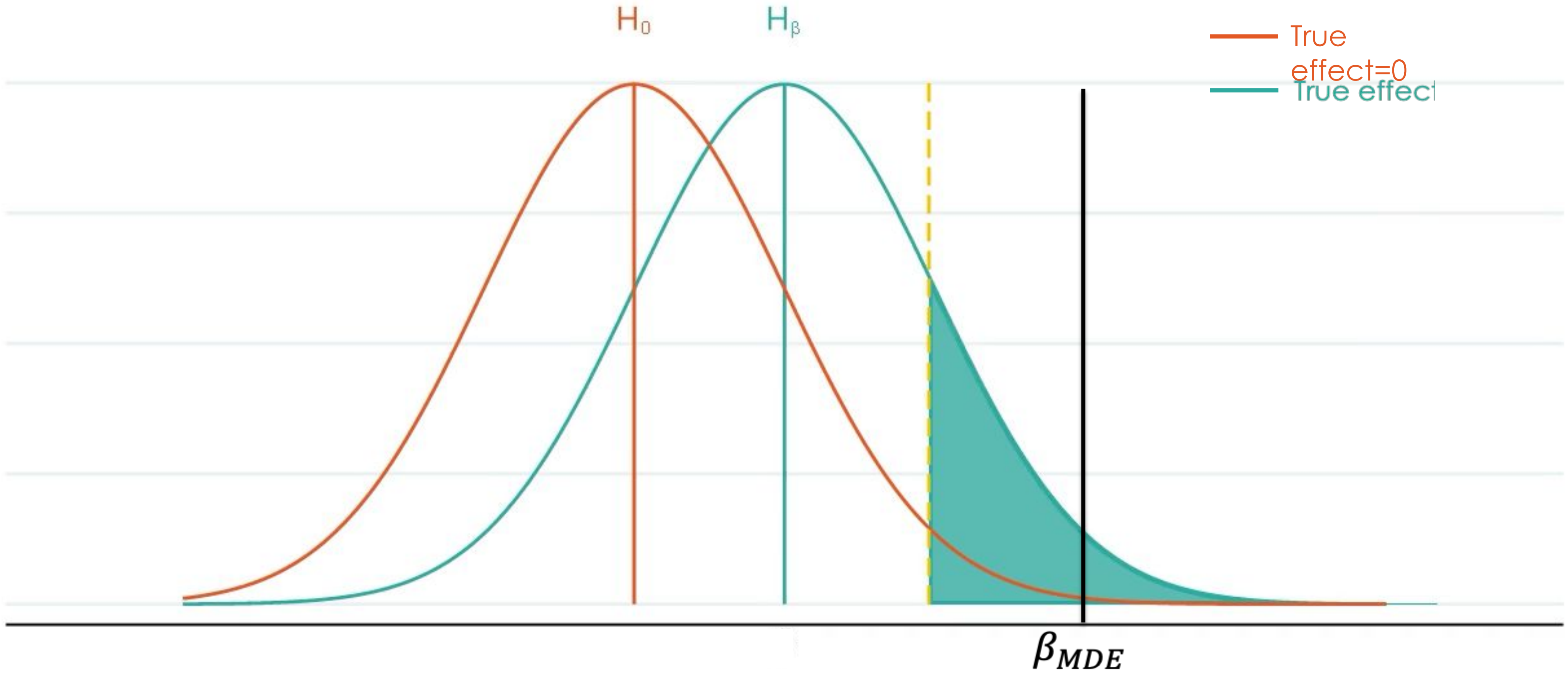
How narrow must the sampling distribution be for there to be 80% of the mass to the right of the critical value given the effect size?

Minimum detectable effect size (MDE)



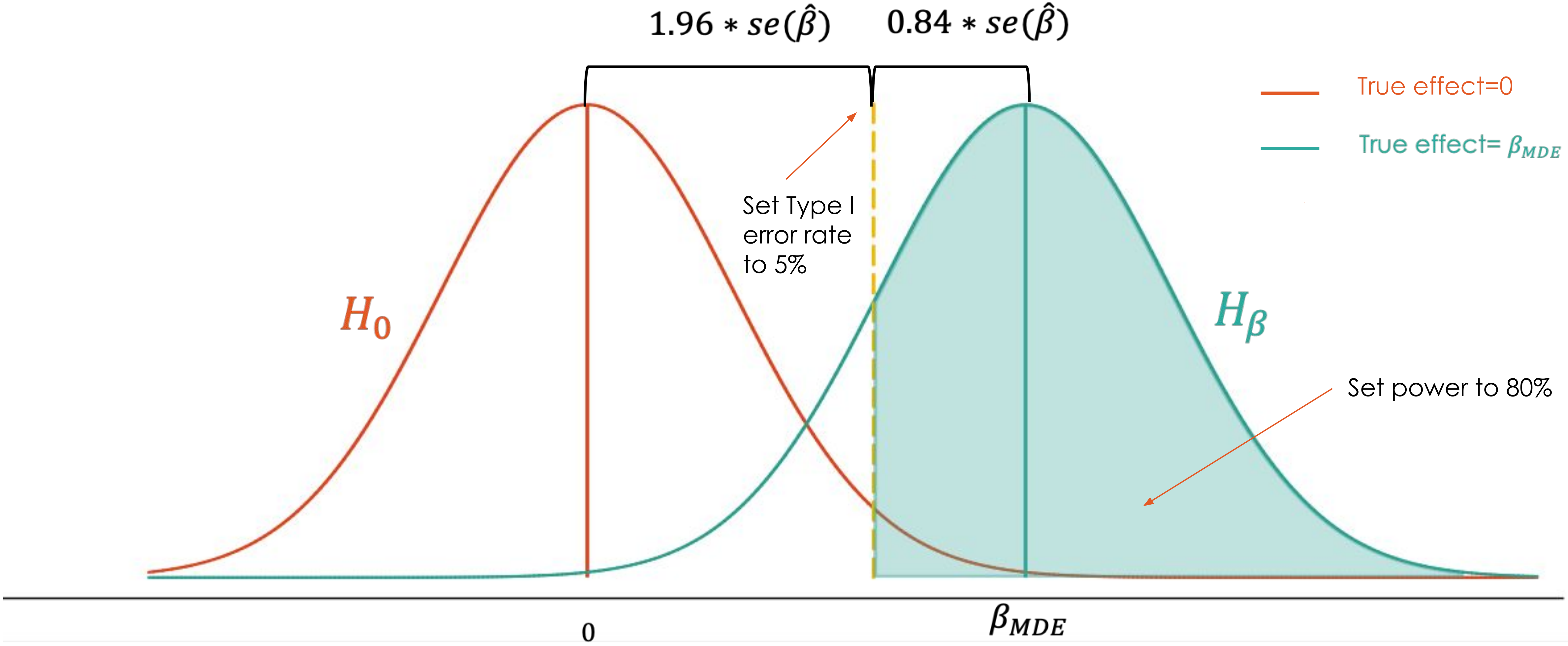
Minimum detectable effect (MDE)

Calculating minimum detectable effect (MDE) for a given sample size



How large must the true effect, β, be for there to be 80% of the mass to the right of the critical value given *N*?

Calculating minimum detectable effect (MDE)

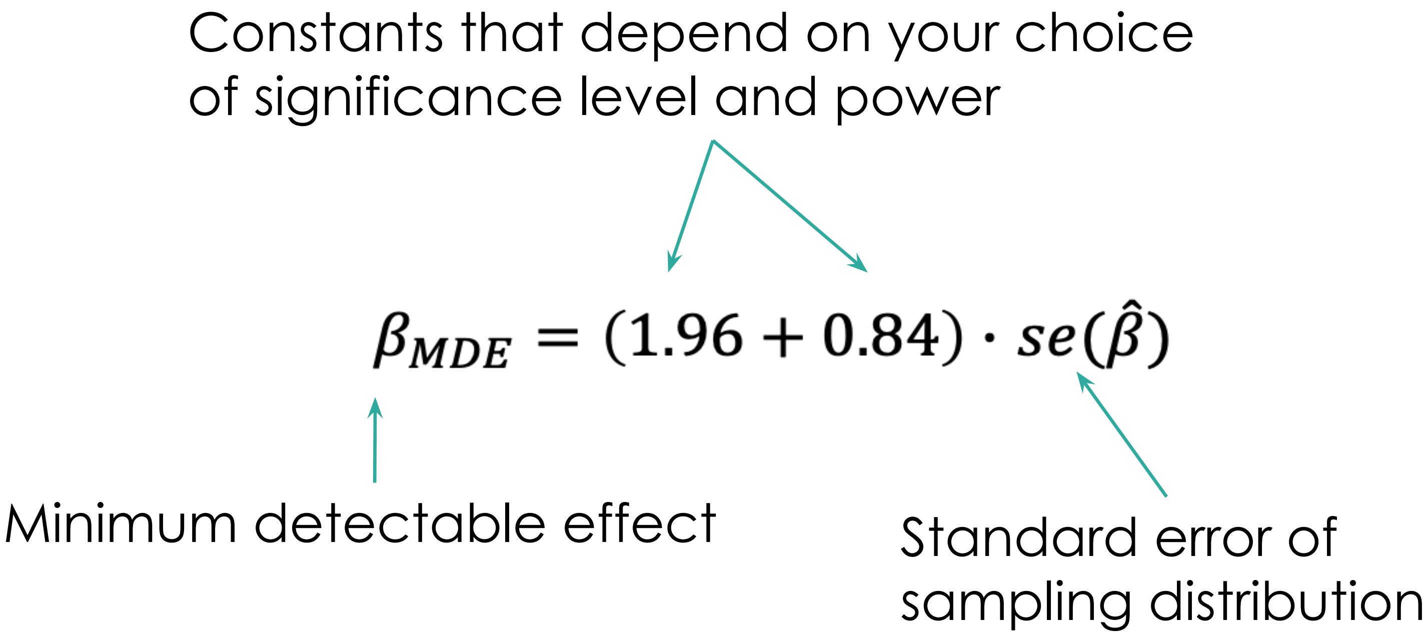


Set Type I error rate to 5%

Set power to 80%

$$\beta_{MDE} = 1.96 \cdot se(\hat{\beta}) + 0.84 \cdot se(\hat{\beta}) = (1.96 + 0.84) \cdot se(\hat{\beta})$$

Calculating the minimum detectable effect size (MDE)



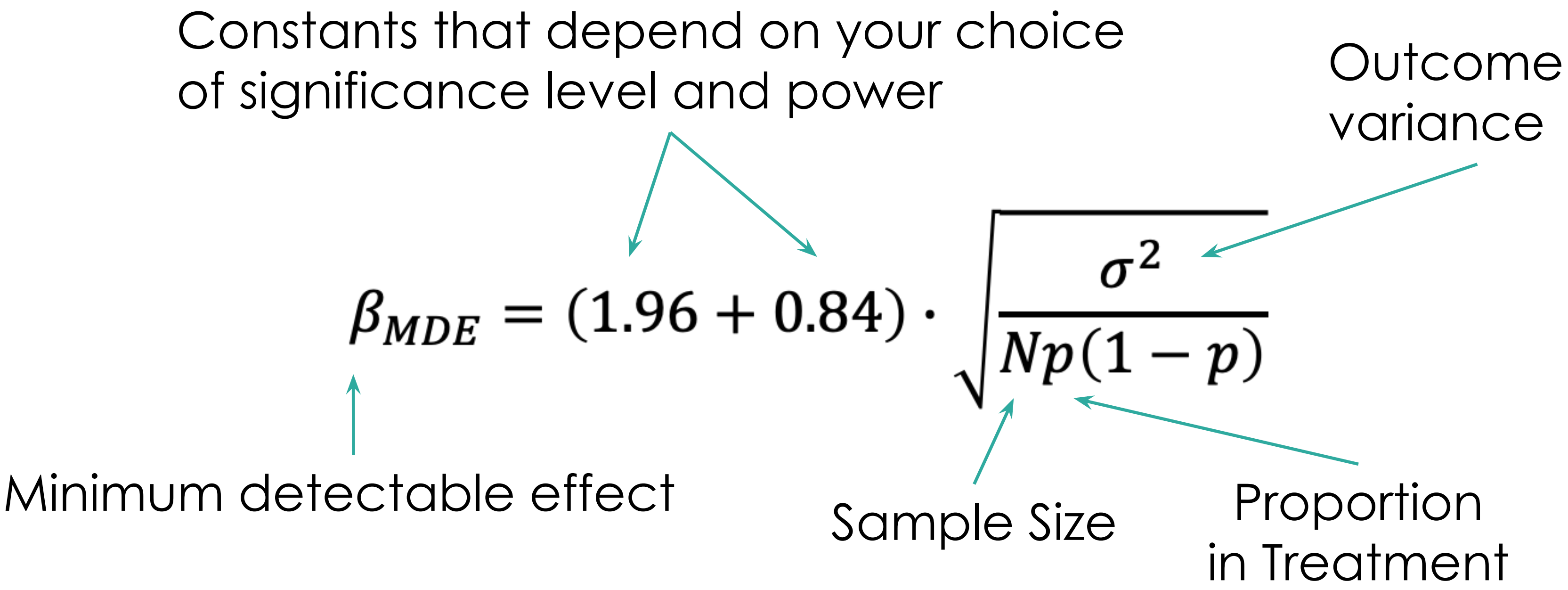
Constants that depend on your choice of significance level and power

$$\beta_{MDE} = (1.96 + 0.84) \cdot se(\hat{\beta})$$

Minimum detectable effect

Standard error of sampling distribution

Calculating the minimum detectable effect size (MDE)



Constants that depend on your choice of significance level and power

Outcome variance

$$\beta_{MDE} = (1.96 + 0.84) \cdot \sqrt{\frac{\sigma^2}{Np(1-p)}}$$

Minimum detectable effect

Sample Size

Proportion in Treatment

The MDE will be smaller with

- Larger sample size *N*
- Smaller outcome variance σ^2
- Even allocation ratio $(p = 0.5)$

For the derivation, see Athey, S., & Imbens, G. W. (2017). The econometrics of randomized experiments. In *Handbook of Economic Field Experiments*.

Outline

I. Motivation

II. Hypothesis testing and statistical power

III. Power calculations

IV. Power in cluster-randomized studies

V. Power calculations in practice

VI. Example (if time)

How does the unit of randomization affect power?

- In practice, we often randomize at units larger than the individual, while still measuring outcomes at the individual-level
  - Schools, classrooms, households, villages
- Challenge: Units within clusters are not independent of one another
  - Students from same school likely to have similar family income, test scores, etc.
  - People within households likely to have similar levels of education, political preferences, etc.
- Impact of clustering on power depends on how "similar" units within a given cluster are (intra-cluster correlation coefficient, ICC, ranging from 0 to 1)

Example: Clustering and power

- Research question: Who will win the next local election in your town?
  - Population consists of 10,000 inhabitants: 2,500 households with 4 people in each
- You have resources to poll 200 people and want to get the best possible estimate of who will win
- Who do you poll?:
  - All four people in 50 households
  - One person in 200 households
  - Somewhere in between

Example: Clustering and power

- Research question: Who will win the next local election in your town?
  - Population consists of 10,000 inhabitants: 2,500 households with 4 people in each
- You have resources to poll 200 people and want to get the best possible estimate of who will win
- Who do you poll?:
  - All four people in 50 households
  - **One person in 200 households**
  - Somewhere in between

High intra-cluster correlation:

Units within clusters are very similar to each other → adding more units within a cluster adds little information about the underlying distribution

Example: Clustering and power

- Research question: Do people prefer strawberry or raspberry flavor?
  - Population consists of 10,000 inhabitants: 2,500 households with 4 people in each
- You have resources to poll 200 people and want to get the best possible estimate of what people prefer
- Who do you poll?:
  - All four people in 50 households
  - One person in 200 households
  - Somewhere in between

Example: Clustering and power

- Research question: Do people prefer strawberry or raspberry flavor?
  - Population consists of 10,000 inhabitants: 2,500 households with 4 people in each
- You have resources to poll 200 people and want to get the best possible estimate of what people prefer
- Who do you poll?:
  - **All four people in ~~50~~ 70 households**
  - One person in 200 households
  - **Somewhere in between**

Low intra-cluster correlation:

Units within clusters are not very similar to each other → adding more units within a cluster or adding new clusters both add information about the underlying distribution

How the intra-cluster correlation affects power

- Samples with high intra-cluster correlation have similar units within clusters
  - Adding additional units from the same clusters adds less new information about the underlying distribution than adding units from new clusters
  - Power increases as new clusters are added but is relatively unaffected when new units within a cluster are added
  - ICC=1: You need as many clusters as you would need units if individually randomized
- Samples with low intra-cluster correlation have more variance within clusters
  - Each cluster resembles the underlying population more closely
  - Power increases similarly whether new clusters or new units within existing clusters are added
  - ICC=0: You need as many units as you would need units if individually randomized

Calculating the minimum detectable effect size in a cluster-randomized design

Intra-cluster correlation coefficient

$$\beta_{MDE} = (1.96 + 0.84) \cdot \sqrt{\frac{\sigma^2}{Jp(1-p)}} \cdot \sqrt{\frac{1 + (m-1) \cdot ICC}{m}}$$

Minimum detectable effect

Cluster size

Number of clusters

The MDE in a clustered RCT will be smaller with:

- More clusters , J
- More observations per cluster, m (if ICC<1)
- NB: Typically, the gain in power from increasing the number of clusters is much larger than increasing the number of units in a cluster

Individually randomized:

$$\beta_{MDE} = (1.96 + 0.84) \cdot \sqrt{\frac{\sigma^2}{Np(1-p)}}$$

Outline

I. Motivation

II. Hypothesis testing and statistical power

III. Power calculations

IV. Power in cluster-randomized studies

V. Power calculations in practice

VI. Example (if time)

How to conduct power calculations in practice

$$\beta_{MDE} = \left(t_{1-\alpha/2} + t_{1-\kappa}\right) \cdot \sqrt{\frac{\sigma^2}{Jp(1-p)}} \cdot \sqrt{\frac{1+(m-1)\cdot ICC}{m}}$$

1. Set desired power (e.g. 80%) and significance level (e.g. 5%)
2. Decide allocation ratio of the sample into treatment and control (you can revisit this later)
3. Set sample size, number of clusters, and cluster size (if applicable) based on the budget, availability, and capacity constraints — adjust the sample size based on expected attrition
4. Estimate variance & ICC from data (often the most challenging step)
5. Back out the MDE for each outcome of interest, subgroup analysis, and comparison across treatment arms — adjust the MDE based on expected compliance
6. Conduct sensitivity analyses, incl. "best case" and "worst case" scenarios
7. Ask: Is the range of MDEs realistic/policy-relevant

$$\beta_{MDE} = (1.96 + 0.84) \cdot \sqrt{\frac{\sigma^2}{Jp(1-p)}} \cdot \sqrt{\frac{1 + (m-1) \cdot ICC}{m}}$$

What can you do to improve the power of a study?

Design levers to improve power I

- Increase the **sample size**
  - Conduct individual-level randomized studies when possible
  - Increase the number of units or clusters
  - Reduce attrition

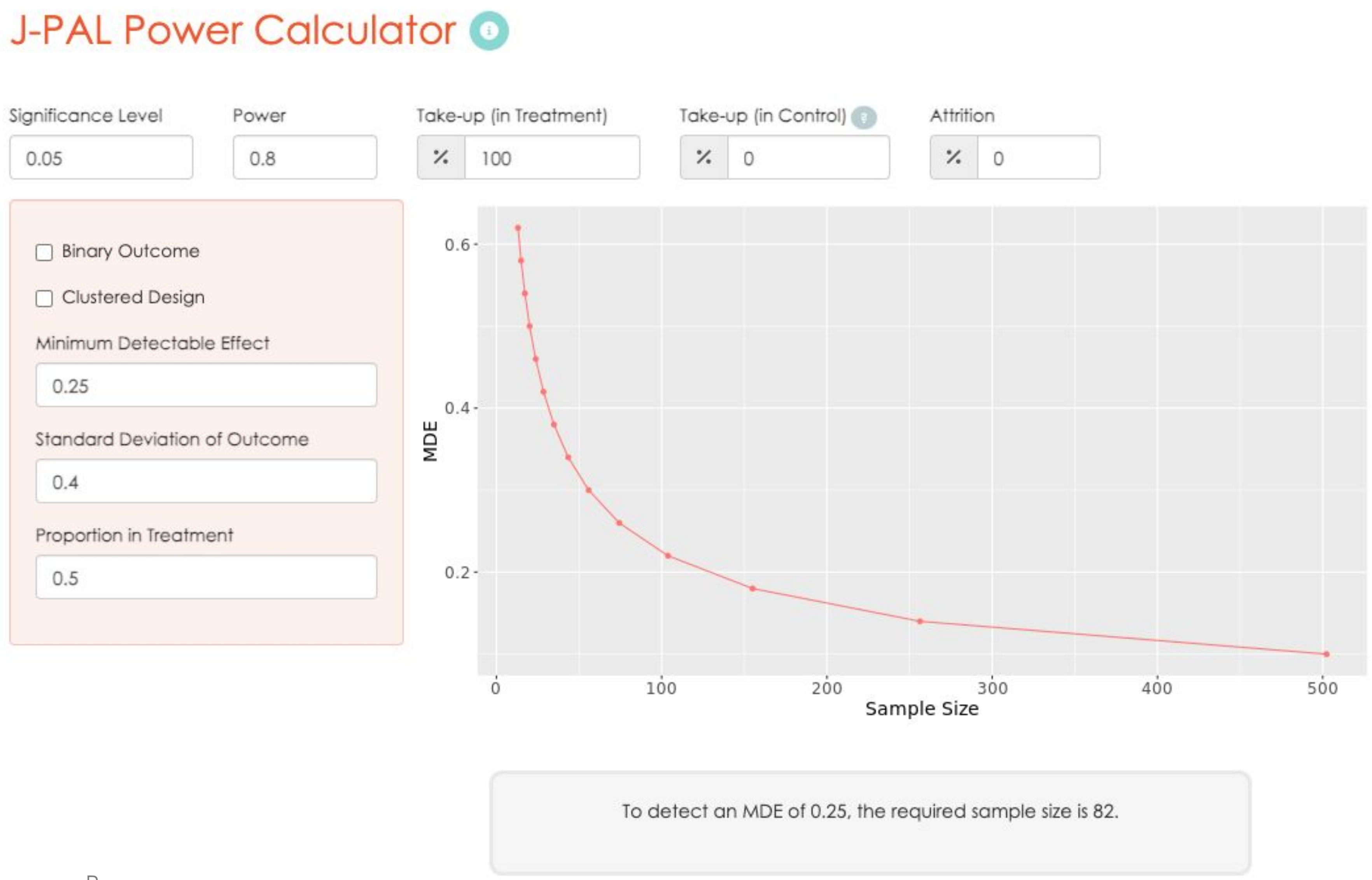
- Increase the **effect size**
  - Increase the intensity of the treatment
  - Increase take-up/compliance
  - Choose an outcome measure that is closely aligned with your TOC
  - (Beware of binary outcomes)

Design levers to improve power II

- Reduce the **outcome variance**
  - Add covariates (especially baseline measure of outcome of interest)
  - Increase data quality/precision
  - Stratify the randomization on important observables

- Reduce the number of **hypotheses you test** (i.e., number of treatment arms, number of subgroups)
  - The study needs to be powered for the smallest MDE among the tests

Preliminary power calculations



J-PAL Power Calculator

Significance Level: 0.05

Power: 0.8

Take-up (in Treatment): % 100

Take-up (in Control): % 0

Attrition: % 0

□ Binary Outcome

□ Clustered Design

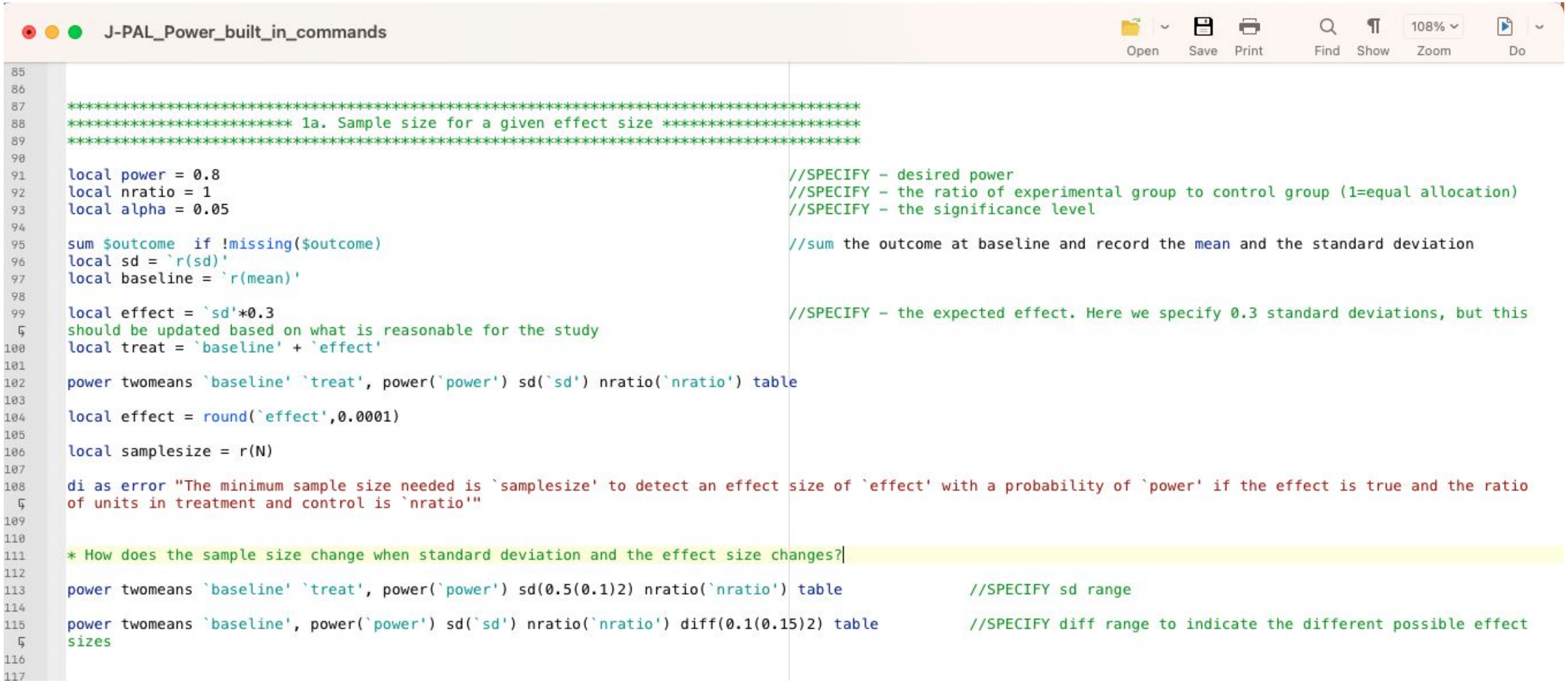
Minimum Detectable Effect: 0.25

Standard Deviation of Outcome: 0.4

Proportion in Treatment: 0.5

To detect an MDE of 0.25, the required sample size is 82.

Final power calculations



```
****************************************************************************************
************************* 1a. Sample size for a given effect size **********************
****************************************************************************************

local power = 0.8                                                               //SPECIFY - desired power
local nratio = 1                                                                //SPECIFY - the ratio of experimental group to control group (1=equal allocation)
local alpha = 0.05                                                              //SPECIFY - the significance level

sum $outcome  if !missing($outcome)                                             //sum the outcome at baseline and record the mean and the standard deviation
local sd = `r(sd)'
local baseline = `r(mean)'

local effect = `sd'*0.3                                                         //SPECIFY - the expected effect. Here we specify 0.3 standard deviations, but this
should be updated based on what is reasonable for the study
local treat = `baseline' + `effect'

power twomeans `baseline' `treat', power(`power') sd(`sd') nratio(`nratio') table

local effect = round(`effect',0.0001)

local samplesize = r(N)

di as error "The minimum sample size needed is `samplesize' to detect an effect size of `effect' with a probability of `power' if the effect is true and the ratio
of units in treatment and control is `nratio'"


* How does the sample size change when standard deviation and the effect size changes?

power twomeans `baseline' `treat', power(`power') sd(0.5(0.1)2) nratio(`nratio') table                    //SPECIFY sd range

power twomeans `baseline', power(`power') sd(`sd') nratio(`nratio') diff(0.1(0.15)2) table                 //SPECIFY diff range to indicate the different possible effect
sizes
```

Practical tips for conducting power calculations

- Perform power calculations early in the design phase (before the program is implemented)

- Don’t panic about the number of assumptions required
  - Power calculations should be considered *guidelines* in the decision of *whether* to carry out the study and provide an *estimate* of how large the sample should be.

- Conduct sensitivity analyses to test how power changes with changes to any critical assumptions
  - Create “best case” scenarios and “worst case” scenarios and evaluate those
  - If the best case scenario MDE is unrealistically high/requires an unrealistically large sample size, consider how to tweak the design to increase power
  - If sufficient power cannot be achieved, an RCT might not be the best way forward

Outline

I. Motivation

II. Hypothesis testing and statistical power

III. Power calculations

IV. Power in cluster-randomized studies

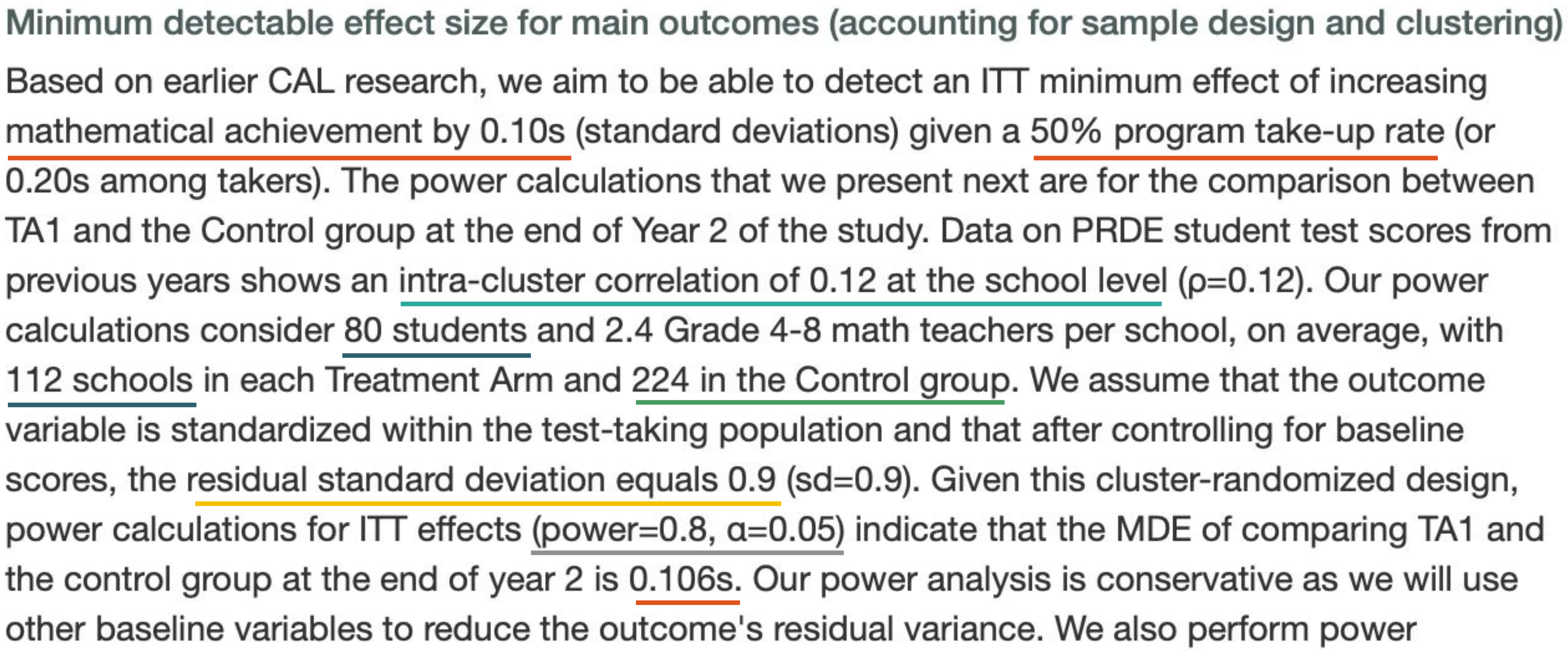
V. Power calculations in practice

VI. **Example** (if time)

Power example

$$\beta_{MDE} = \left(t_{1-\alpha/2} + t_{1-\kappa}\right) \cdot \sqrt{\frac{\sigma^2}{Jp(1-p)}} \cdot \sqrt{\frac{1+(m-1)\cdot ICC}{m}}$$

Bobonis, Gustavo J. et al. 2022. "Adopting Computer Assisted Learning (CAL) at Scale: Training and Supporting Teachers and Families to Use CAL Technologies in Puerto Rico Public Schools." AEA RCT Registry. May 30. https://doi.org/10.1257/rct.7720-1.1



Minimum detectable effect size for main outcomes (accounting for sample design and clustering)

Based on earlier CAL research, we aim to be able to detect an ITT minimum effect of increasing mathematical achievement by 0.10s (standard deviations) given a 50% program take-up rate (or 0.20s among takers). The power calculations that we present next are for the comparison between TA1 and the Control group at the end of Year 2 of the study. Data on PRDE student test scores from previous years shows an intra-cluster correlation of 0.12 at the school level (ρ=0.12). Our power calculations consider 80 students and 2.4 Grade 4-8 math teachers per school, on average, with 112 schools in each Treatment Arm and 224 in the Control group. We assume that the outcome variable is standardized within the test-taking population and that after controlling for baseline scores, the residual standard deviation equals 0.9 (sd=0.9). Given this cluster-randomized design, power calculations for ITT effects (power=0.8, α=0.05) indicate that the MDE of comparing TA1 and the control group at the end of year 2 is 0.106s. Our power analysis is conservative as we will use other baseline variables to reduce the outcome's residual variance. We also perform power

Resources for understanding power

- Power guides:
  - Power Calculations (J-PAL)
  - Quick Guide to Power Calculations (J-PAL)
  - Six Rules of Thumb for Power (J-PAL)
  - Ten things to know about power (EGAP)
  - Power calculations in practice handout (J-PAL)

- Data sources for estimating variance, ICC, etc:
  - J-PAL/IPA Dataverse
  - World Bank Microdata Library and LSMS data
  - IPUMS or DHS data (large health and population household surveys)
  - National statistics, administrative data, etc.

Resources for calculating power

STATA

- Sample code on conducting power in Stata and R (J-PAL)
- Power calculations in STATA (World Bank)
- Power by simulation in STATA (World Bank)
- power and clustersampsi commands (Stata)

R

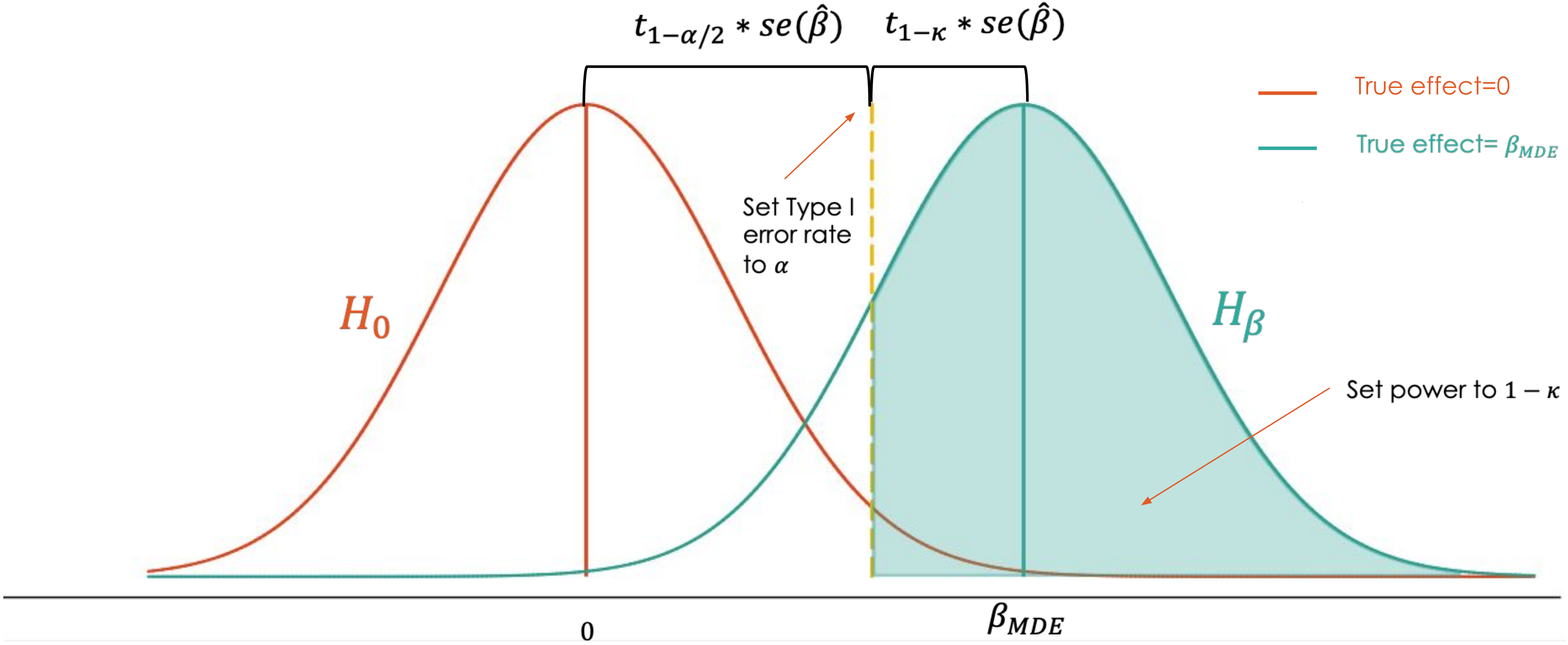
- There are many ways to conduct power calculations in R: one way is to use the pwrcalc package (github)
- Simulation in R (EGAP)

References

- Athey, S., and Imbens, G. W. (2017). The econometrics of randomized experiments. *Handbook of Economic Field Experiments,* 73-140.
- Banerjee, A., Cole, S., Duflo, E., and Linden, L. (2007). Remedying Education: Evidence from Two Randomized Experiments in India. *The Quarterly Journal of Economics,* 122(3), 1235–1264.
- Bruhn, M. and McKenzie, D. (2008). In Pursuit of Balance: Randomization is Practice in Development Field Experiments, Policy Research Working Paper #4752.
- Coppock, A., Gelman, A., and Hill, J. (2007). Sample size and power calculations. *Data analysis using regression and multilevel/hierarchical models,* 437-455.
- Duflo, Esther, Glennerster, R., and Kremer, M. (2008). *Using Randomization in Development Economics research: A Toolkit*. Handbook of Development Economics vol. 4, 3895-3962.
- Glennerster, R., and Takavarasha, K. (2013). *Running Randomized Evaluations: A Practical Guide,* Princeton University Press, Princeton, NJ.
- McConnell, B., and Vera-Hernández, M. (2015). Going beyond simple sample size calculations: a practitioner's guide. IFS Working Papers No. W15/17.

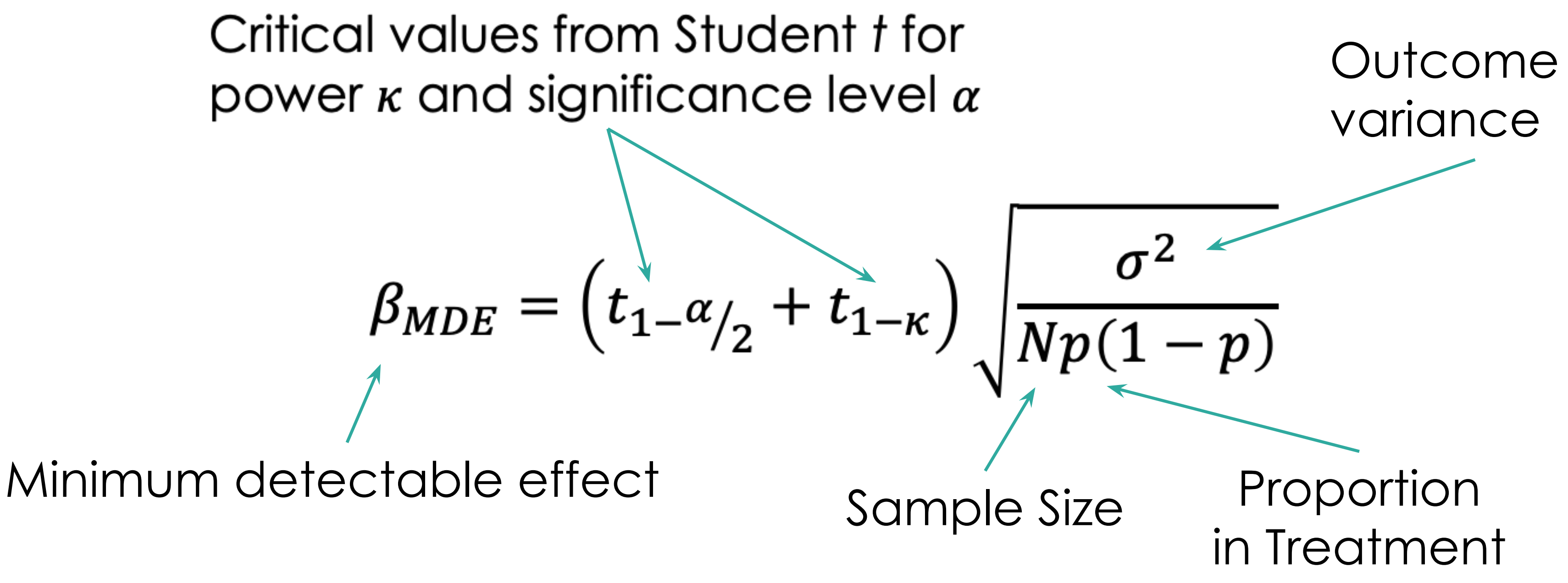
Appendix

Calculating minimum detectable effect (MDE)


$t_{1-\alpha/2} * se(\hat{\beta})$
$t_{1-\kappa} * se(\hat{\beta})$
True effect=0
True effect= β_{MDE}
Set Type I error rate to α
H_0
H_β
Set power to $1-\kappa$
0
β_{MDE}

$$\beta_{MDE} = t_{1-\alpha/2} * se(\hat{\beta}) + t_{1-\kappa} * se(\hat{\beta}) = \left(t_{1-\alpha/_2} + t_{1-\kappa}\right) se(\hat{\beta})$$

Calculating the minimum detectable effect size



Critical values from Student *t* for power κ and significance level α

Outcome variance

$$\beta_{MDE} = \left(t_{1-\alpha/2} + t_{1-\kappa}\right)\sqrt{\frac{\sigma^2}{Np(1-p)}}$$

Minimum detectable effect

Sample Size

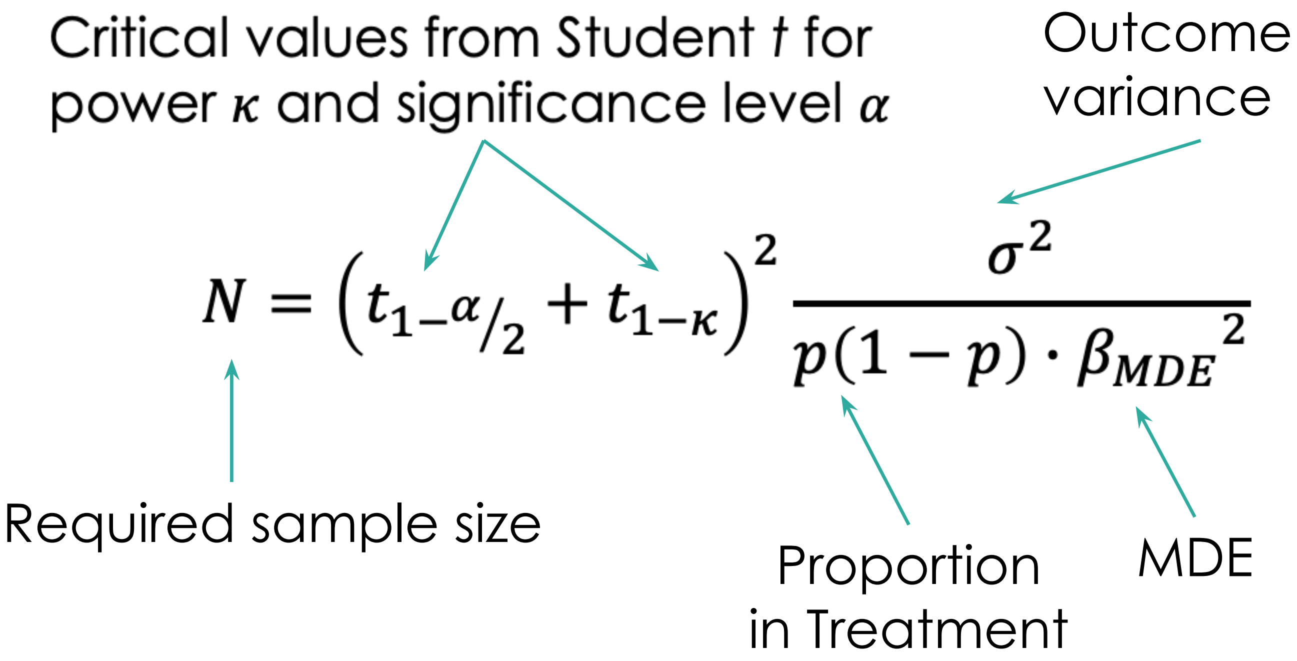
Proportion in Treatment

The MDE will be smaller with

- Larger sample size *N*
- Smaller outcome variance σ^2
- Even allocation ratio ($p = 0.5$)

For the derivation, see Athey, S., & Imbens, G. W. (2017). The econometrics of randomized experiments. In *Handbook of Economic Field Experiments*

Calculating the required sample size



Critical values from Student *t* for power κ and significance level α

Outcome variance

$$N = \left(t_{1-\alpha/_{2}} + t_{1-\kappa}\right)^{2} \frac{\sigma^{2}}{p(1-p)\cdot \beta_{MDE}{}^{2}}$$

Required sample size

Proportion in Treatment

MDE

The required N will be smaller with

- Larger MDE
- Smaller outcome variance σ^2
- Even allocation ratio ($p = 0.5$)

For the derivation, see Athey, S., & Imbens, G. W. (2017). The econometrics of randomized experiments. In *Handbook of Economic Field Experiments*

Power calculations step by step: Calculate sample size

$$N = \left(t_{1-\alpha/2} + t_{1-\kappa}\right)^2 \frac{\sigma^2}{p(1-p)\cdot {\beta_{MDE}}^2}$$

1. Set desired power (e.g. 80%) and significance level (e.g. 5%)
2. Decide allocation ratio of the sample into treatment and control (you can revisit this later)
3. Set MDE based on past studies and/or policy relevance and cost effectiveness — adjust MDE based on expected compliance
4. Estimate variance & ICC (if applicable) from data
5. Back out the sample size — if calculating number of clusters, specify cluster size, and vice versa
6. Conduct sensitivity analyses, incl. "best case" and "worst case" scenarios
7. Ask: Is the range of sample sizes realistic

Calculating the minimum detectable effect size in a cluster-randomized design

Intra-cluster correlation coefficient

$$\beta_{MDE} = \left(t_{1-{\alpha}/{2}} + t_{1-\kappa}\right) \cdot \sqrt{\frac{\sigma^2}{Jp(1-p)}} \cdot \sqrt{\frac{1+(m-1)\cdot ICC}{m}}$$

Minimum detectable effect

Cluster size

Number of clusters

The MDE in a clustered RCT will be smaller with:

- More clusters , J
- More observations per cluster, m (if ICC<1)
- NB: Typically, the gain in power from increasing the number of clusters is much larger than increasing the number of units in a cluster

Reuse and citation

To reference this lecture, please cite as:

J-PAL. “Lecture: Mechanics of Power.” Abdul Latif Jameel Poverty Action Lab. 2024. Cambridge, MA