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. Ethical Considerations for Randomized Evaluations
7. Threats & Analysis
8. Randomized Evaluation from Start to Finish
9. Applying & Using Evidence
10. The Generalizability Framework

Power tracks

- **Essentials of Sample Size and Power (75 minutes)**: 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 (90 minutes)**: 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 Type I and Type II errors
- Understand technically how the power of a study is derived, how it is calculated, and what components of a study affect its power
- Will be convinced of 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. Determinants of power

V. Practical tips

Outline

I. **Motivation**

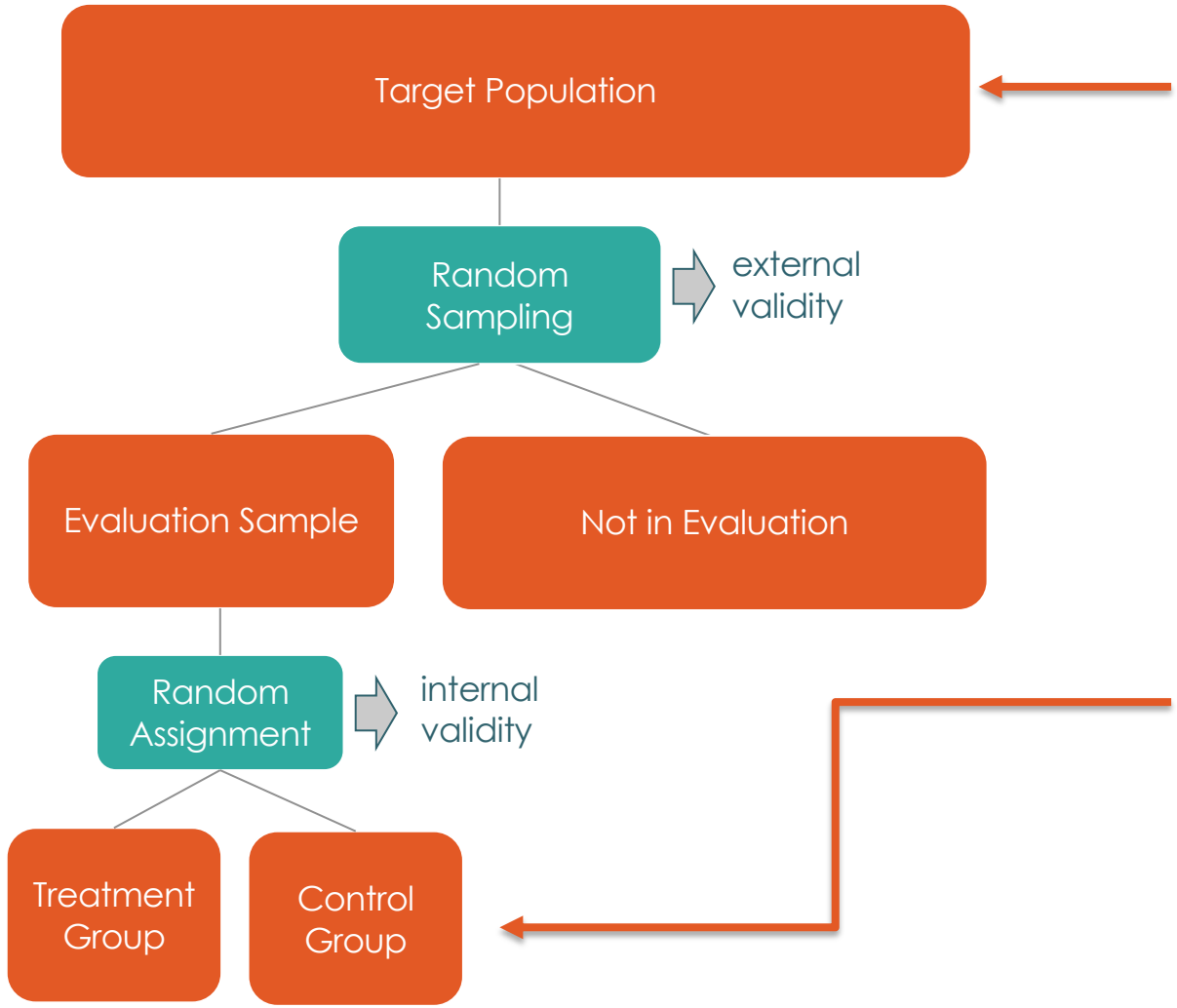
II. Hypothesis testing and statistical power

III. Power calculations

IV. Determinants of power

V. Practical tips

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

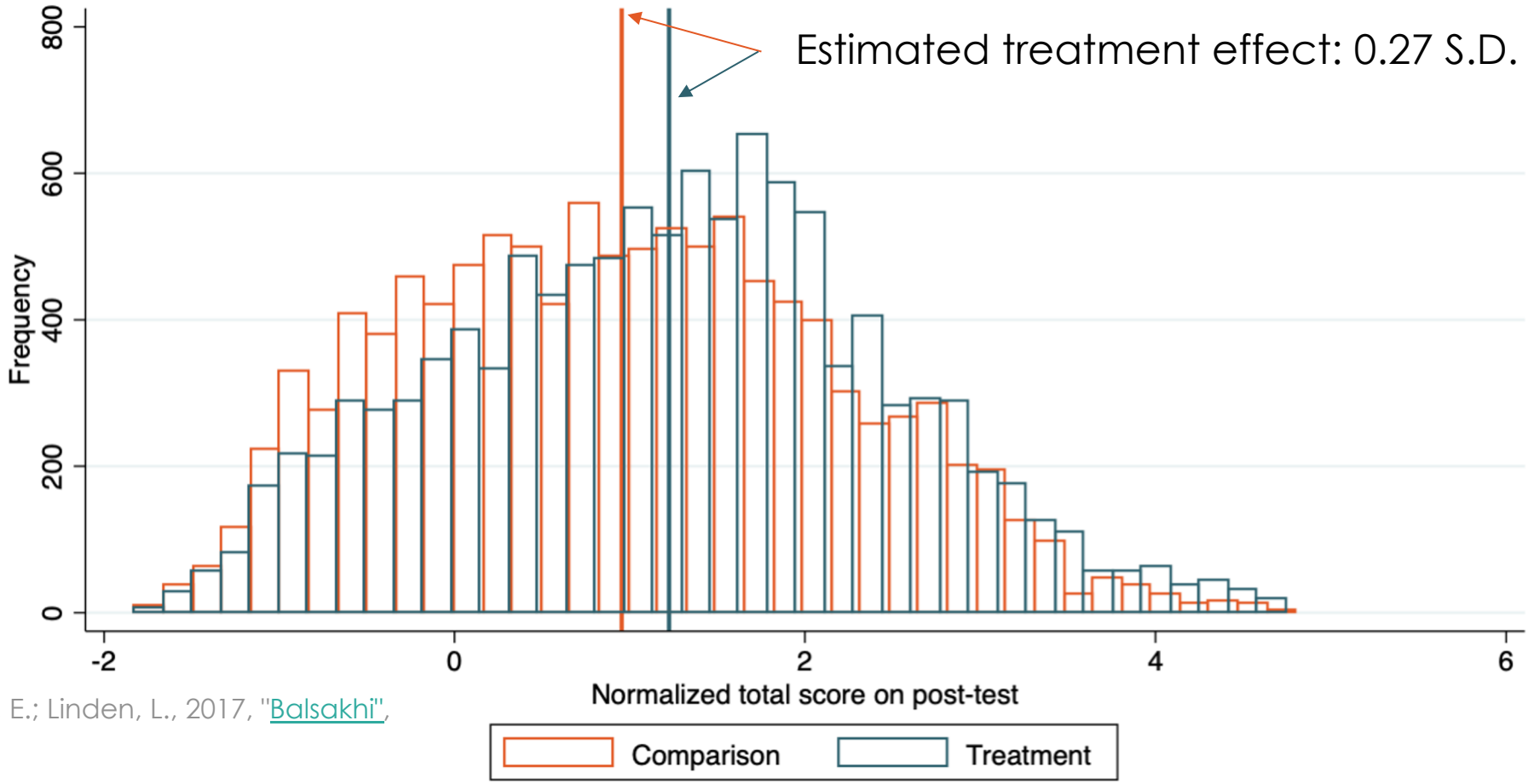
Estimated treatment effect ($\hat{\beta}$): the sample difference in the outcome between the treatment and comparison group

- 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: Balsahki tutoring program

Study: Balsahki remedial tutoring program in India

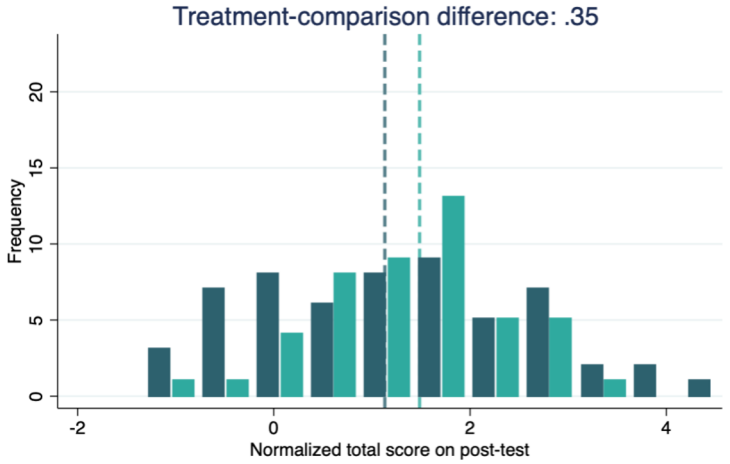
Sample size: More than 23,000 students

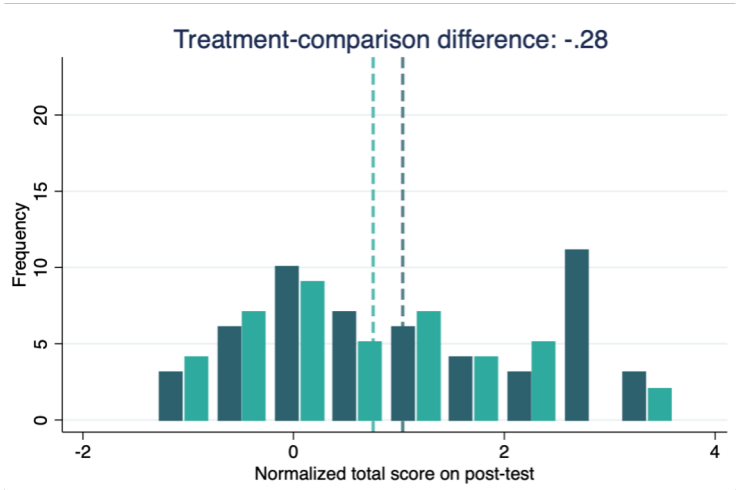


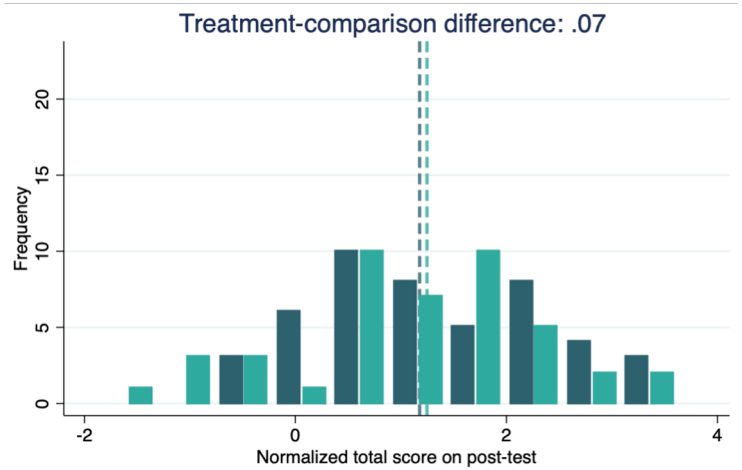
Source: Banerjee, A., Cole, S.; Duflo, E.; Linden, L., 2017, "Balsakhi", Harvard Dataverse

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







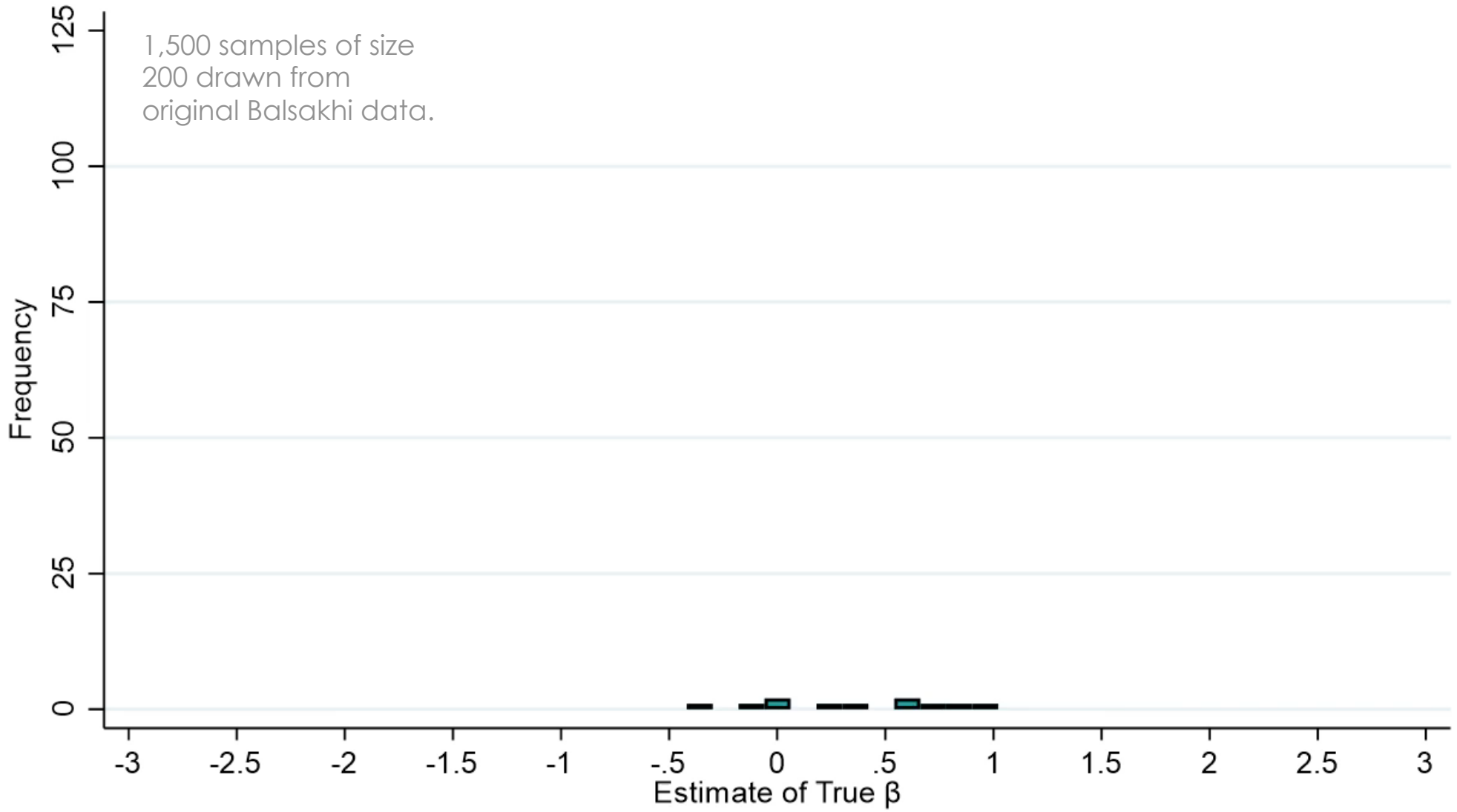
Comparison

Treatment

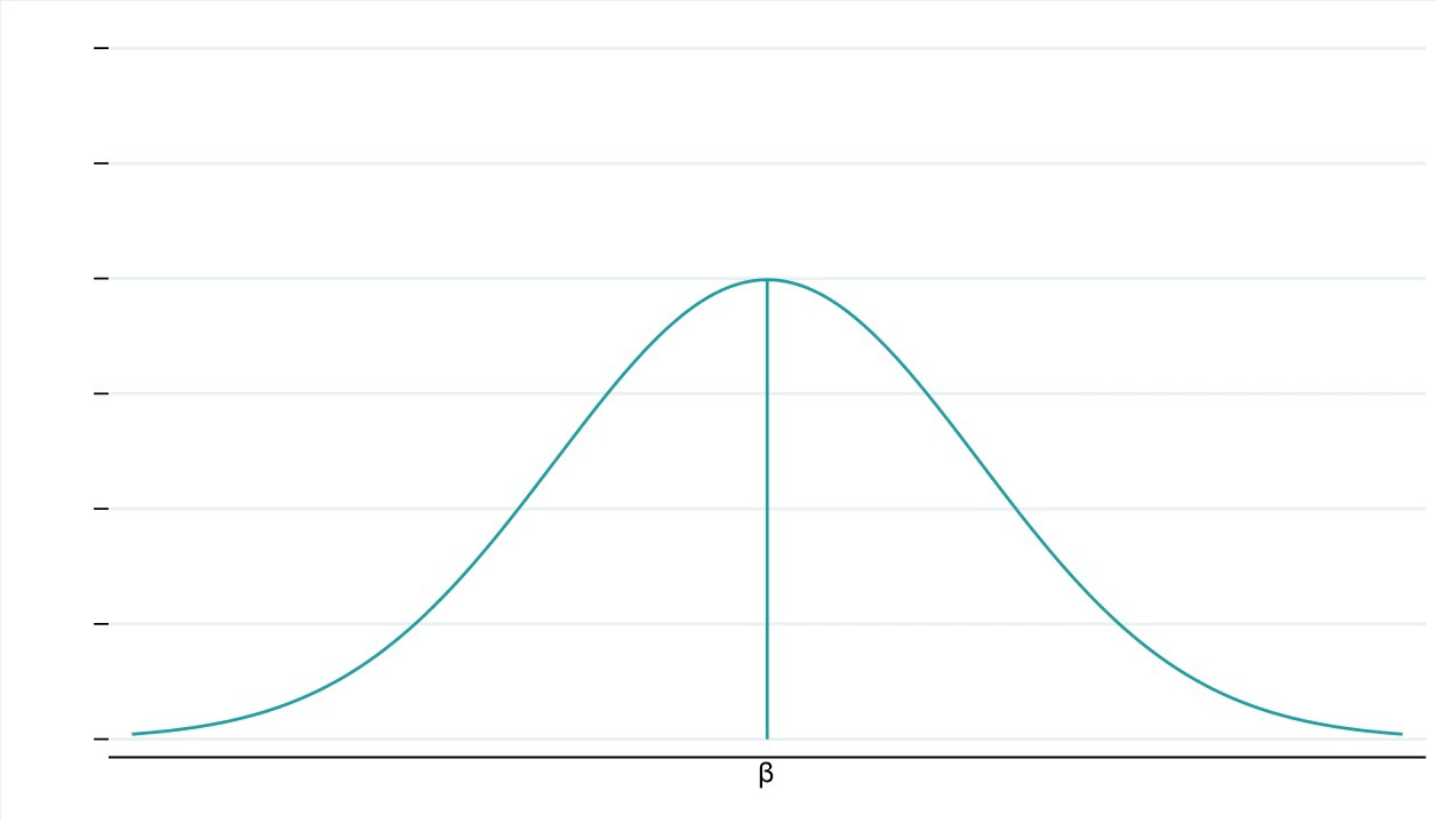
Samples of size 200 drawn from original Balsakhi data.

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

Many samples: a *sampling distribution* of estimates



The larger the sample size, the more narrow the sampling distribution…

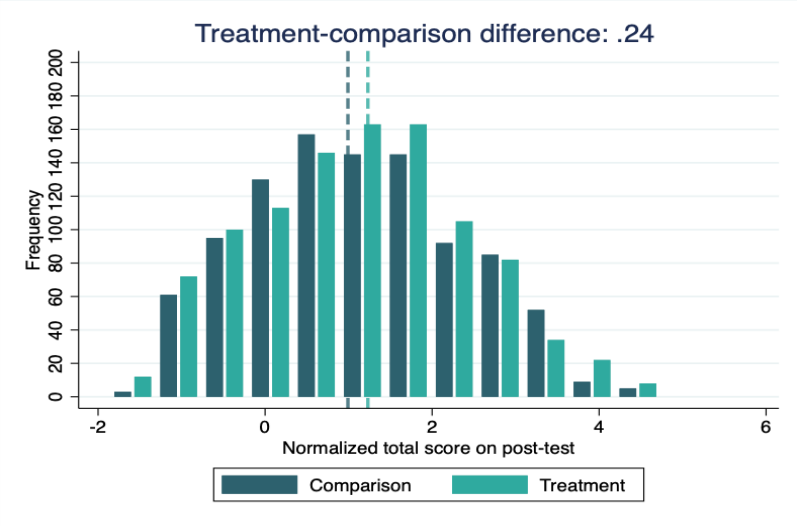


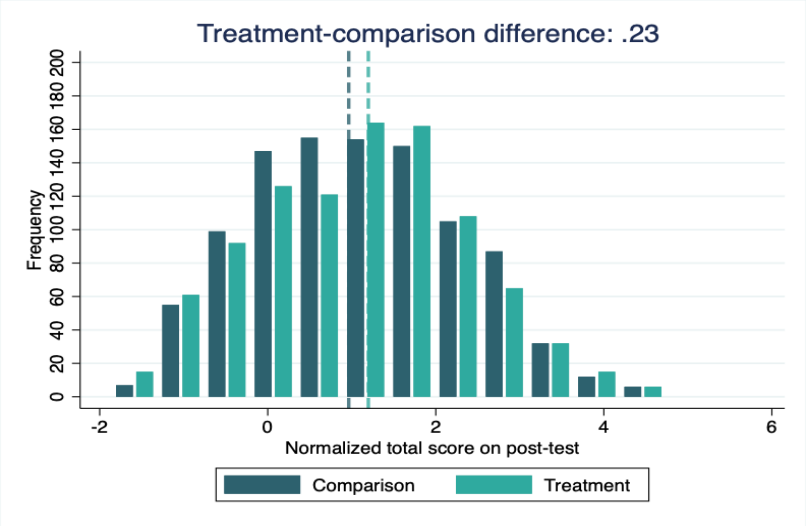
Central limit theorem: Normal distribution

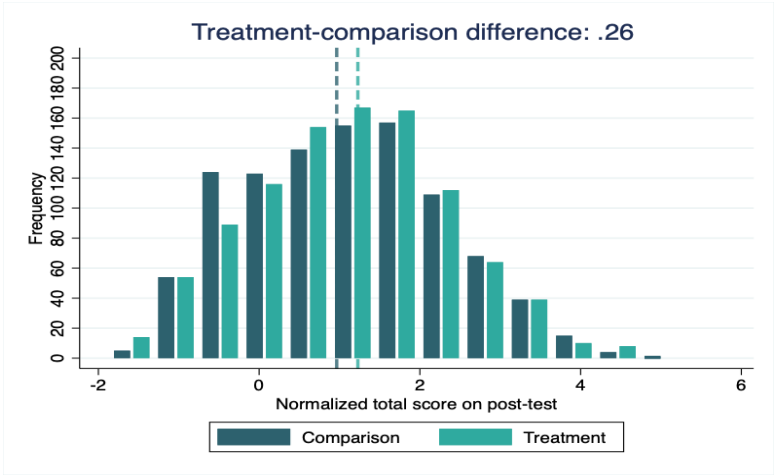
Randomization: Centered around the true treatment effect, β

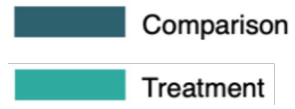
… and the more likely the estimated treatment effect is close to the true treatment effect

Larger samples lead to less random variation in treatment effects





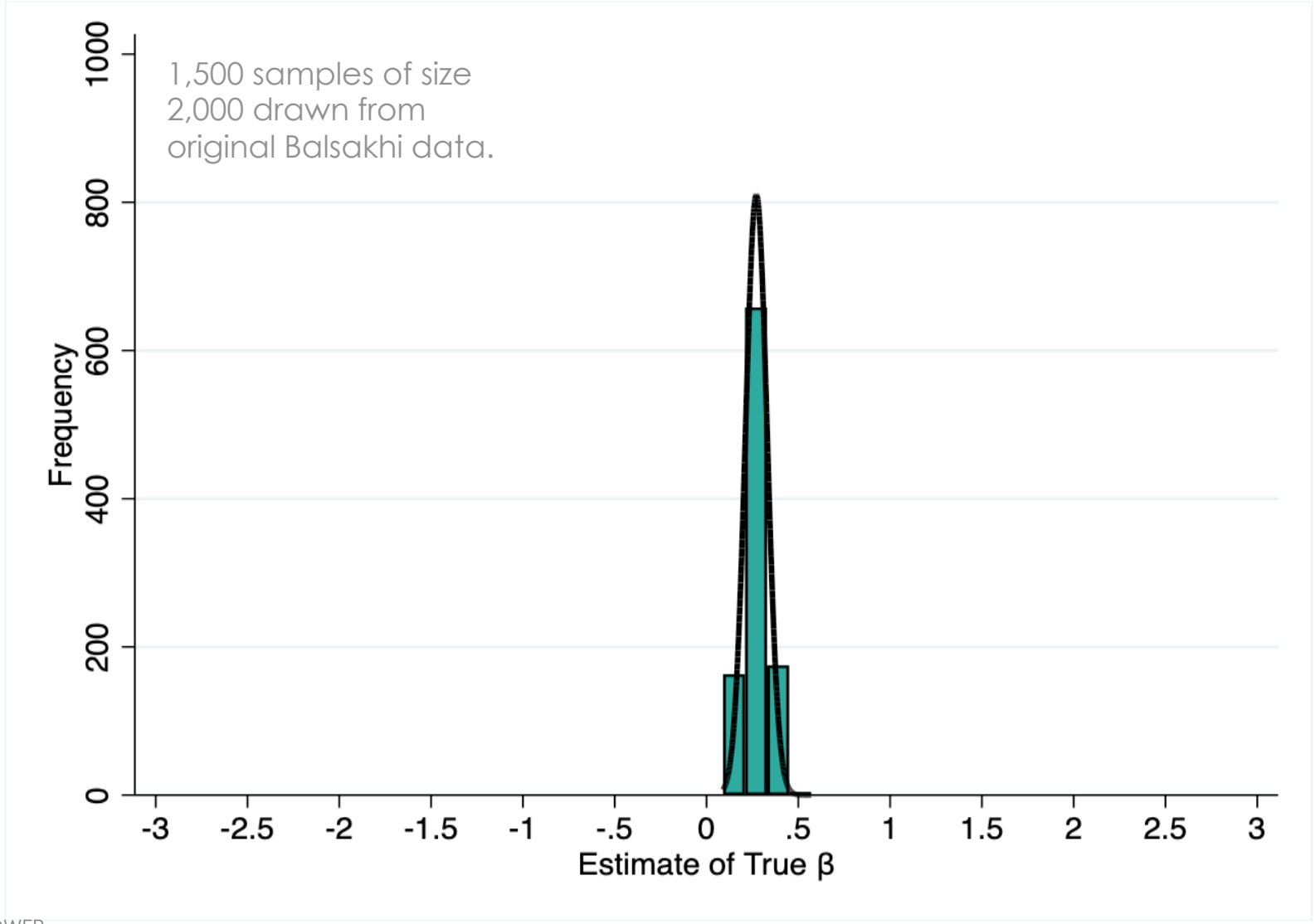




Samples of size 2,000 drawn from original Balsakhi data.

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



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. Determinants of power

V. Practical tips

Hypothesis testing

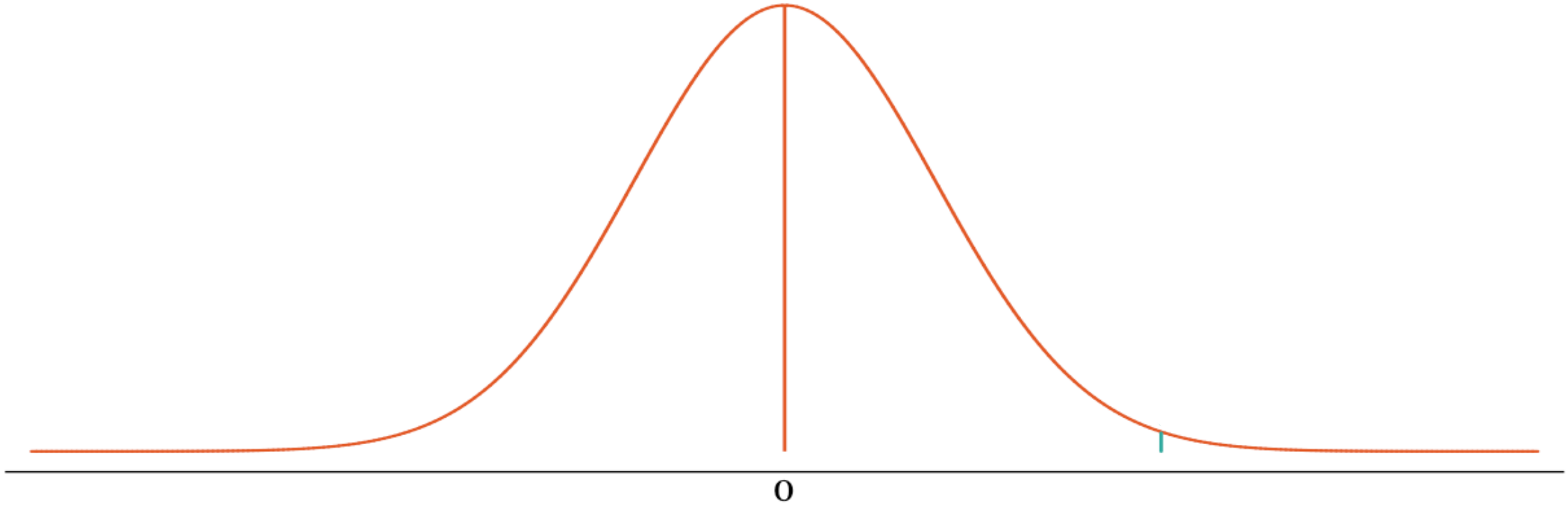
- Researchers and policymakers want to know "Is my program effective?"
  - i.e., did it change the outcome of interest?

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

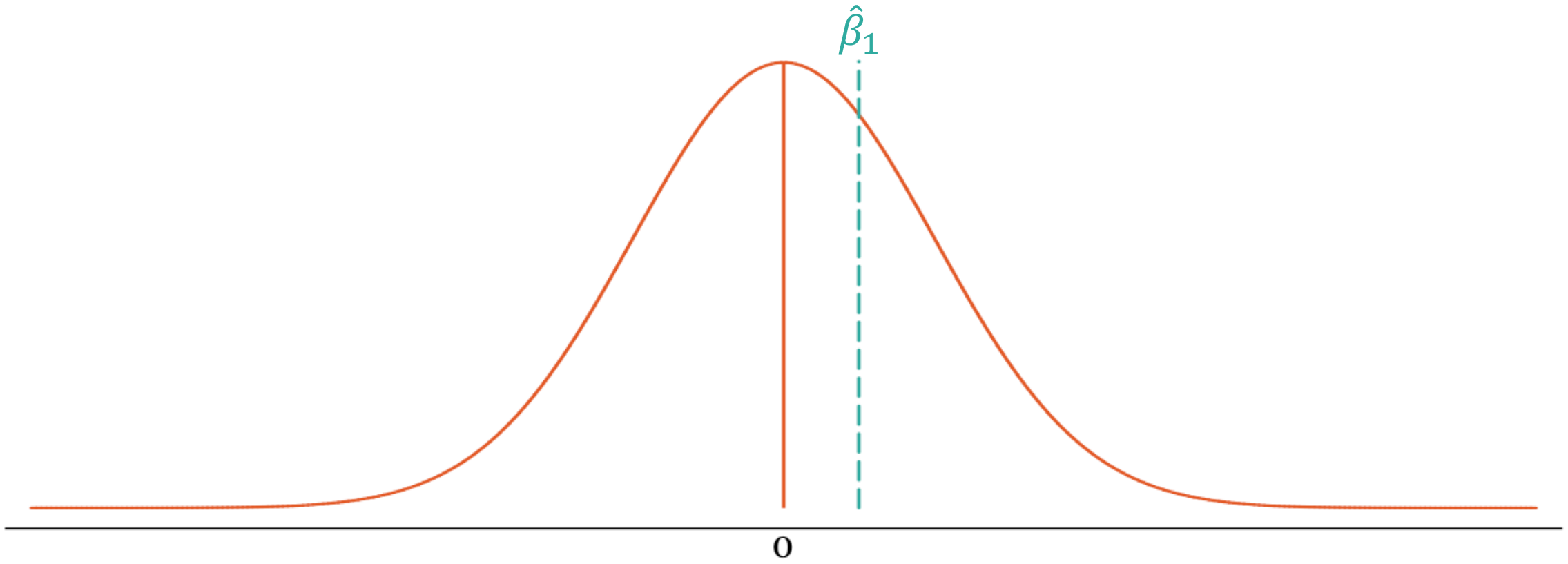
Null hypothesis:

Assume that the true treatment effect is zero



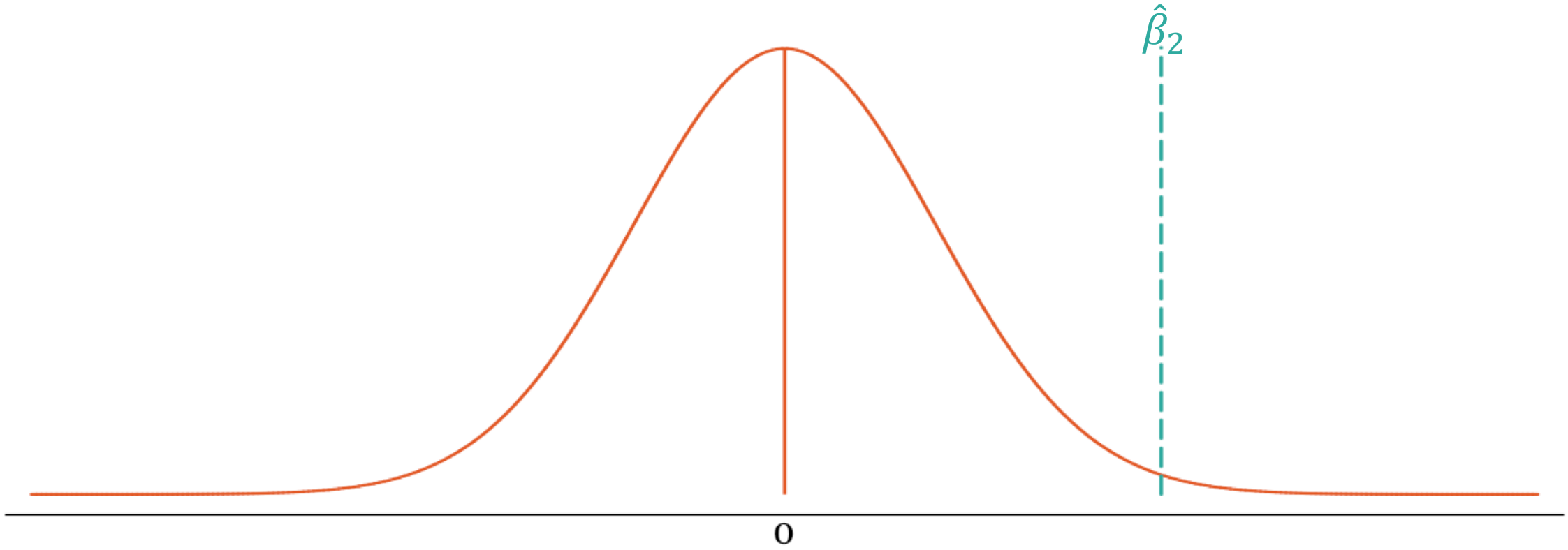
Ask "How likely is it that we would observe the treatment effect estimate, $\hat{\beta}$, if the true effect were zero?"

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



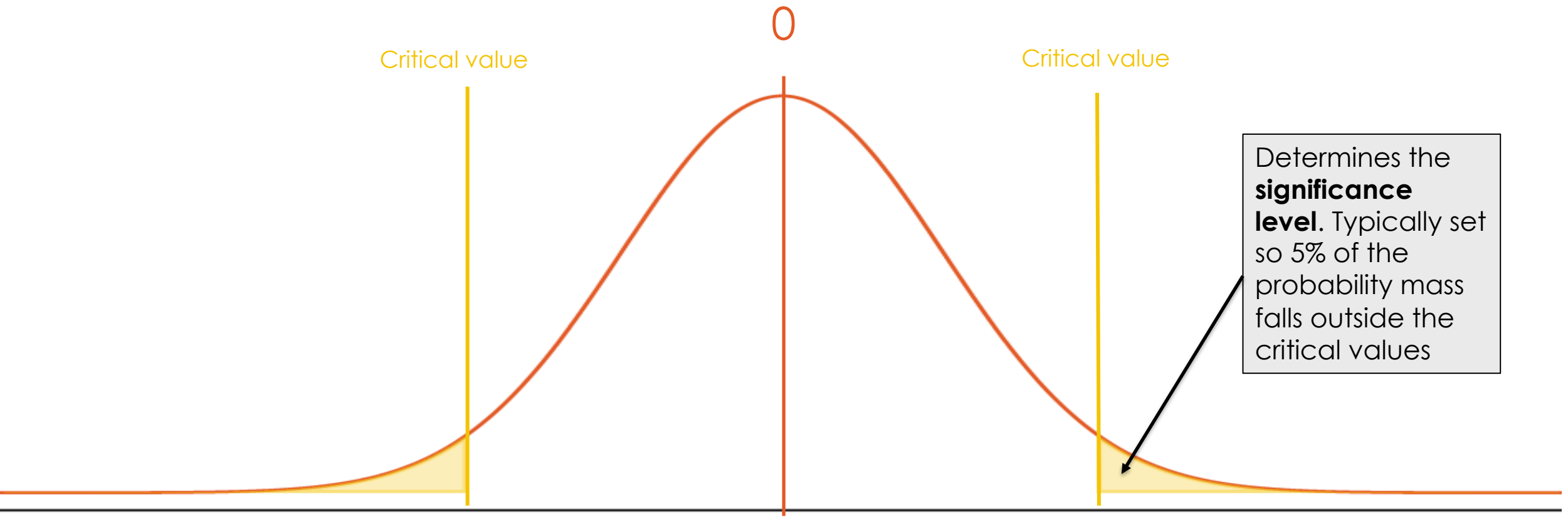
It is rather likely to observe estimates as large as $\hat{\beta}_1$ if the true effect were actually zero

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



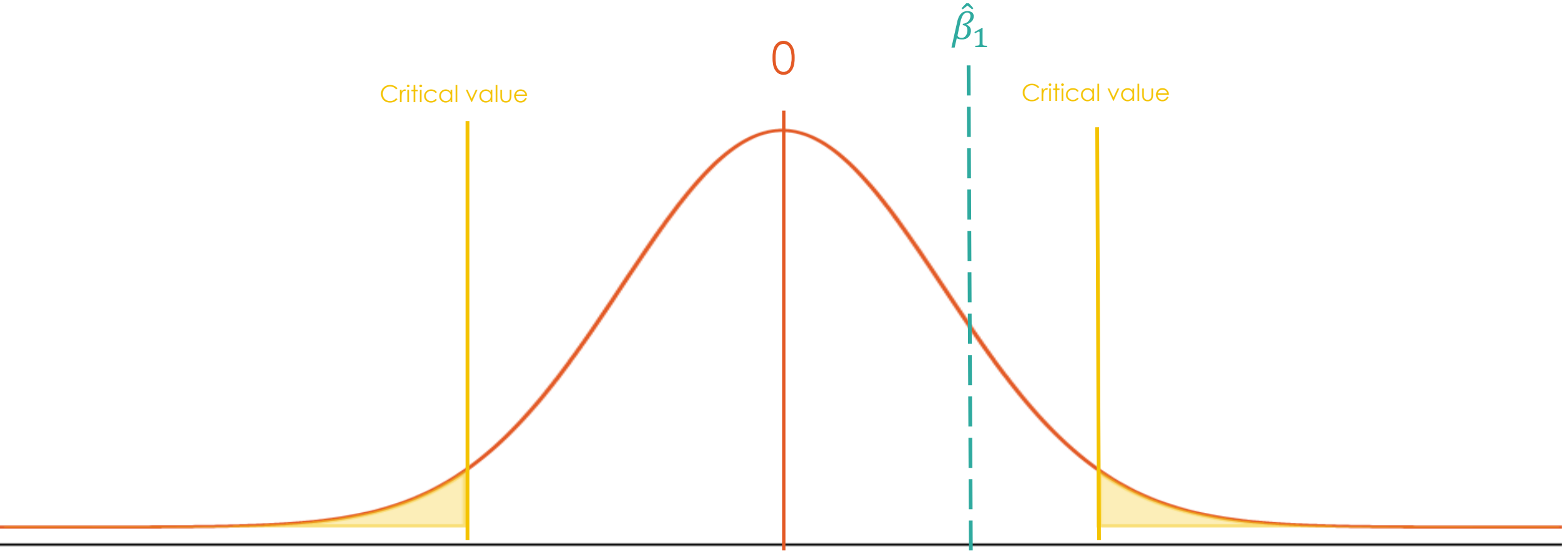
It is rather unlikely to observe estimates as large as $\hat{\beta}_2$ if the true effect were actually zero

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



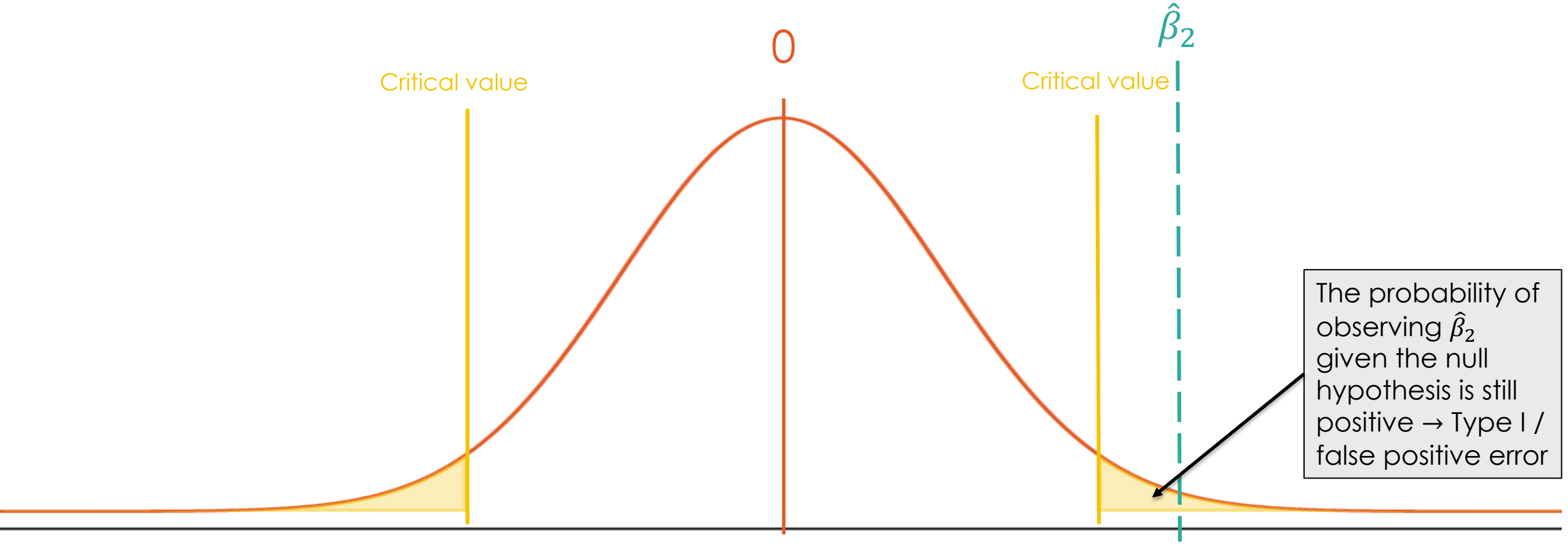
If $\hat{\beta}$ falls outside the critical values, we reject the null hypothesis

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



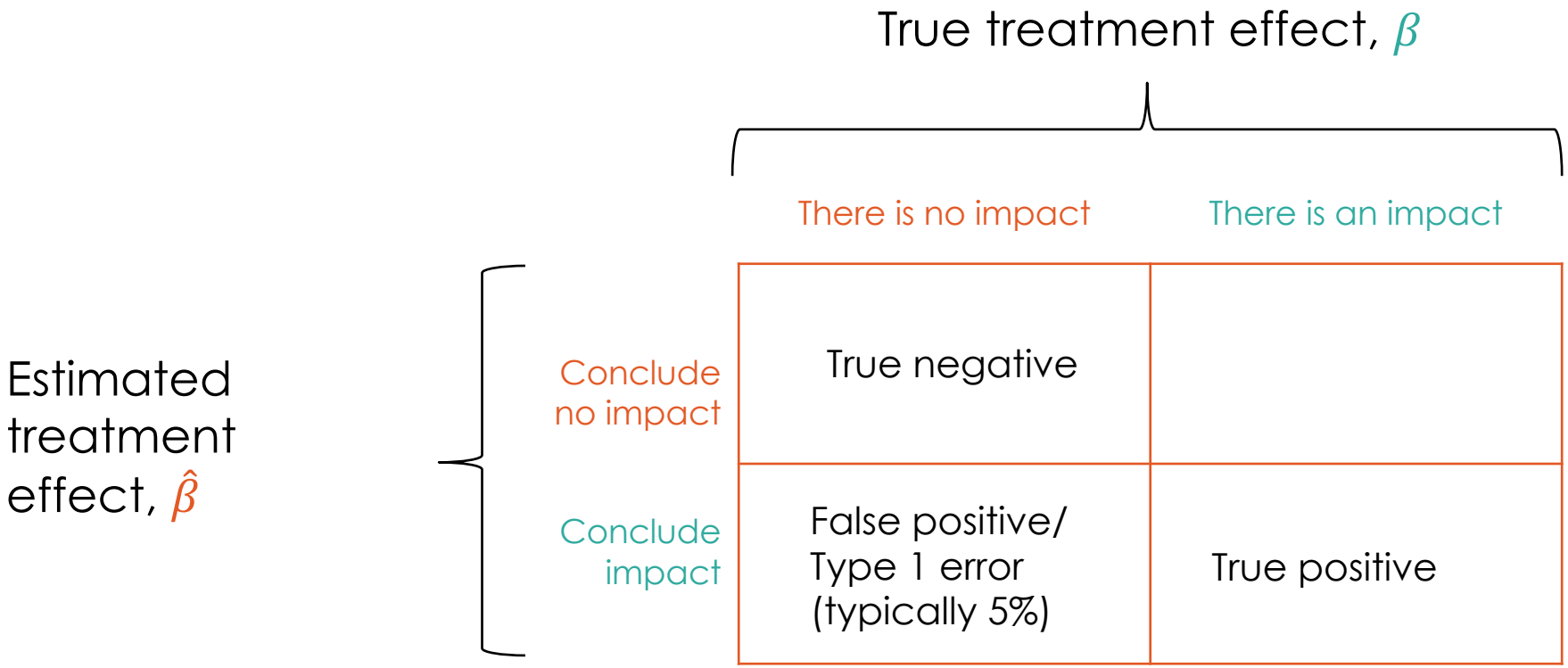
We do <u>not</u> reject the null hypothesis → "$\hat{\beta}_1$ is <u>not</u> statistically significantly different from zero at the 5% level"

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



We do reject the null hypothesis → "$\hat{\beta}_2$ is statistically significantly different from zero at the 5% level"

Evaluation results vs. underlying reality



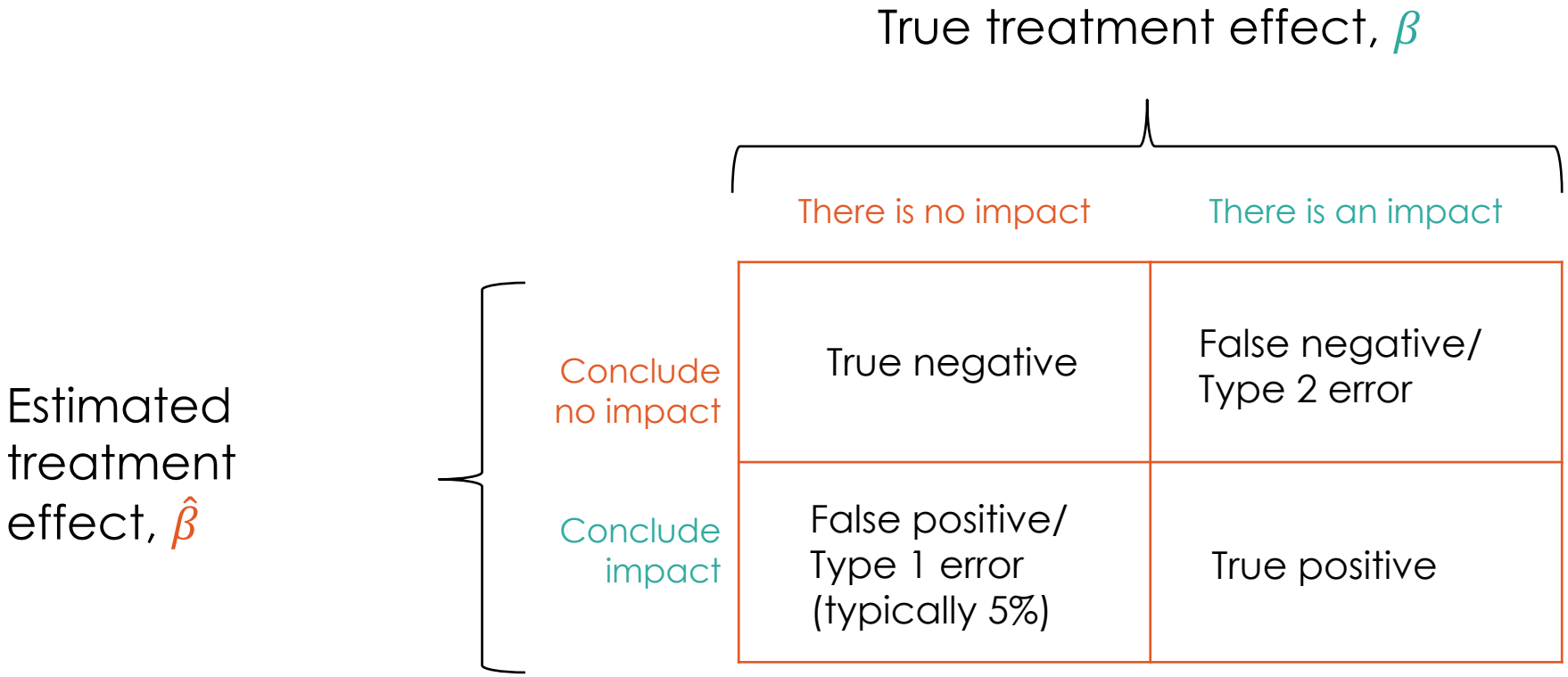
Type I error (false positive)

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

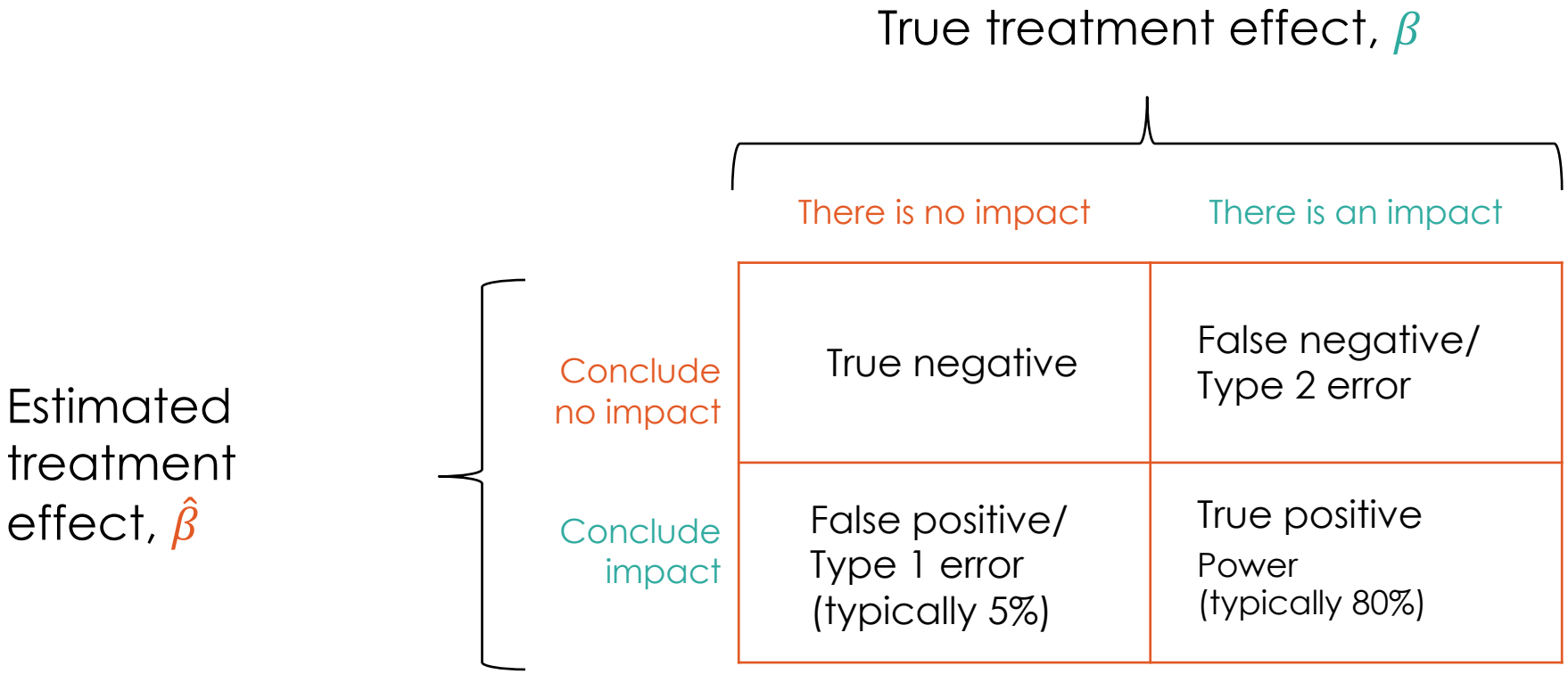


Type II error (false negative)

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?

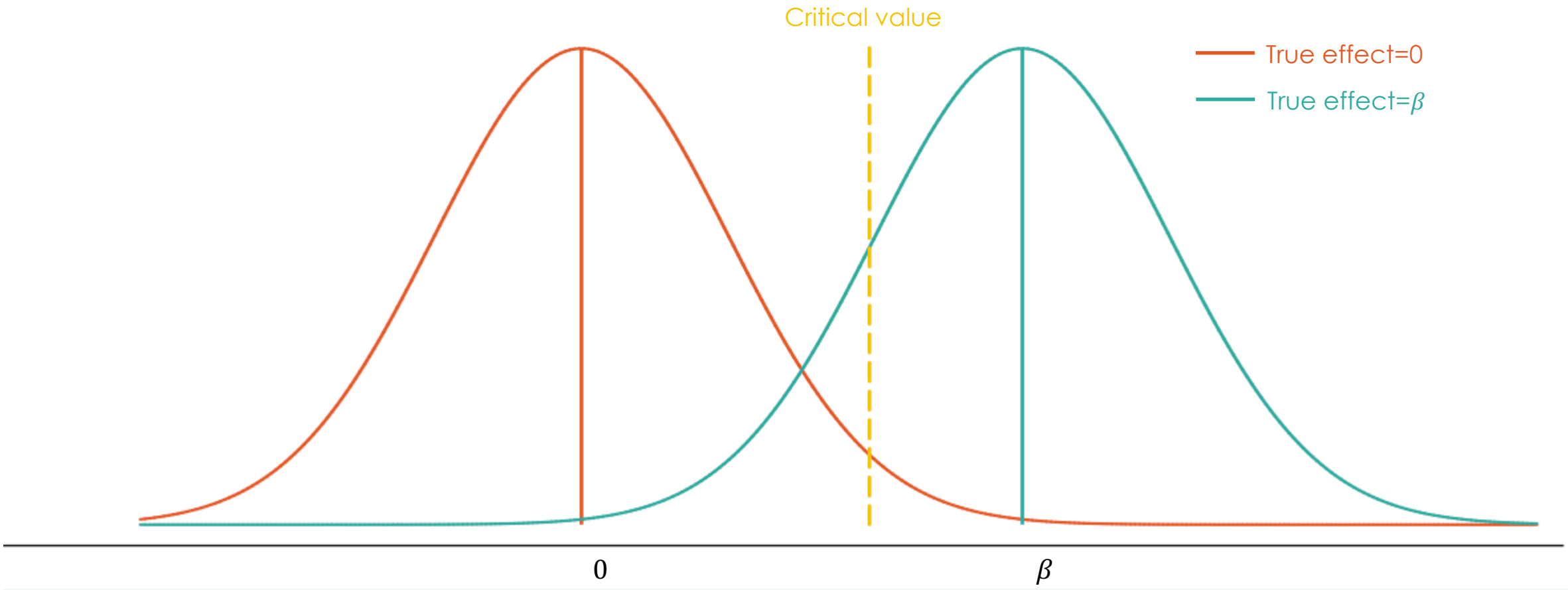
Evaluation results vs. underlying reality



Statistical power (true positive)

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

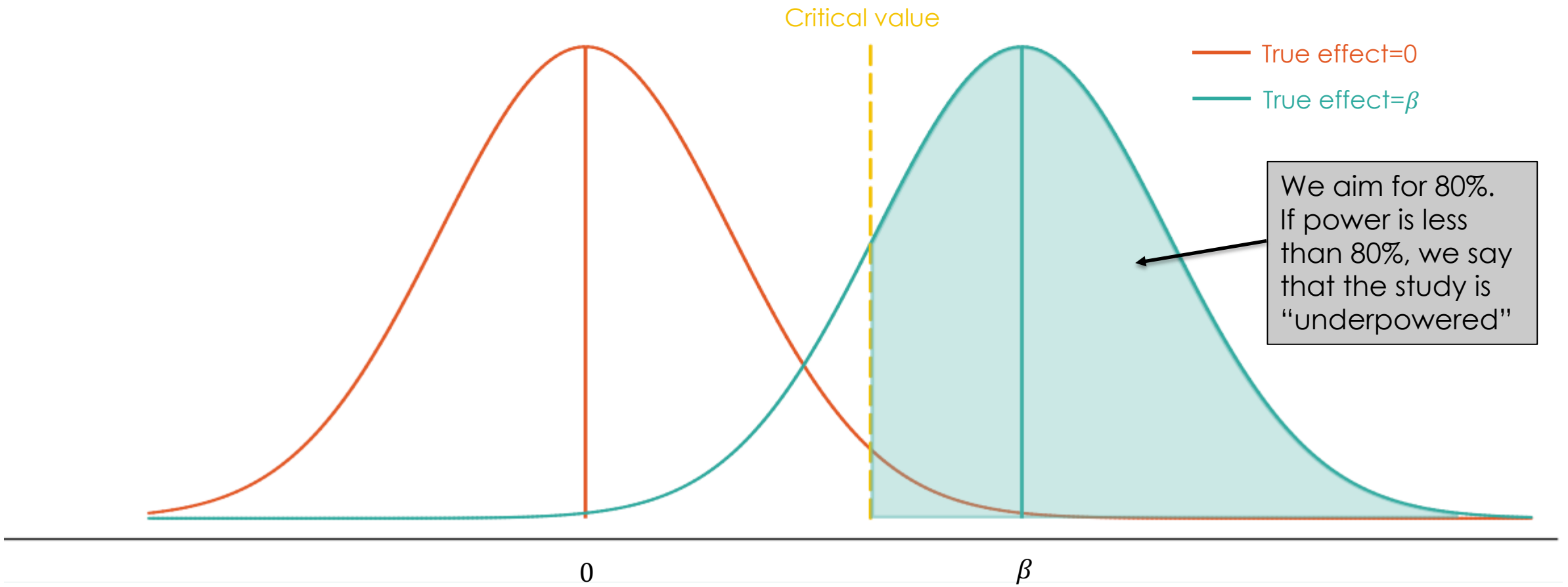
Introducing the alternative hypothesis $\beta \neq 0$



Null hypothesis: Sampling distribution centered around zero

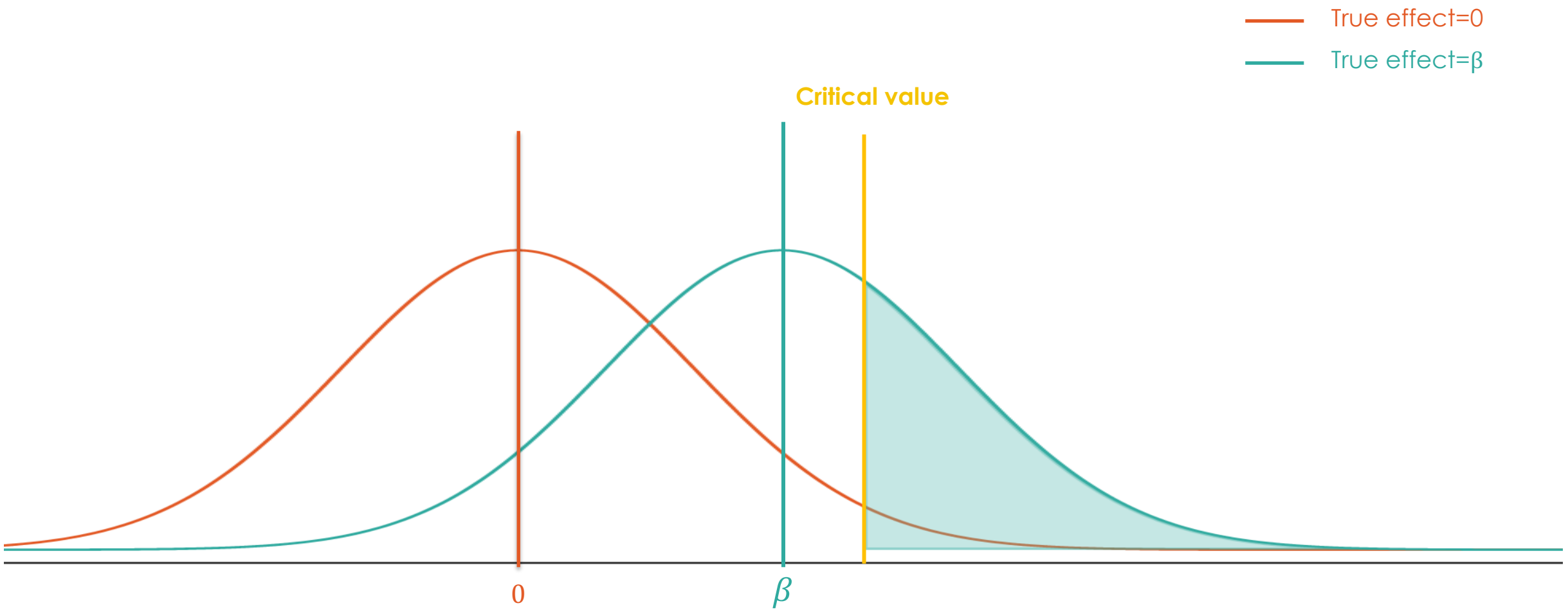
Alternative hypothesis: Same distribution centered around $\beta \neq 0$

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



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

Example of an underpowered study



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

Recap: Type I error, Type II error, and power

True treatment effect, β

Estimated treatment effect, $\hat{\beta}$	There is no impact	There is an impact
Conclude no impact	True negative	False negative/ Type 2 error
Conclude impact	False positive/ Type 1 error (typically 5%)	True positive/ Power (typically 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

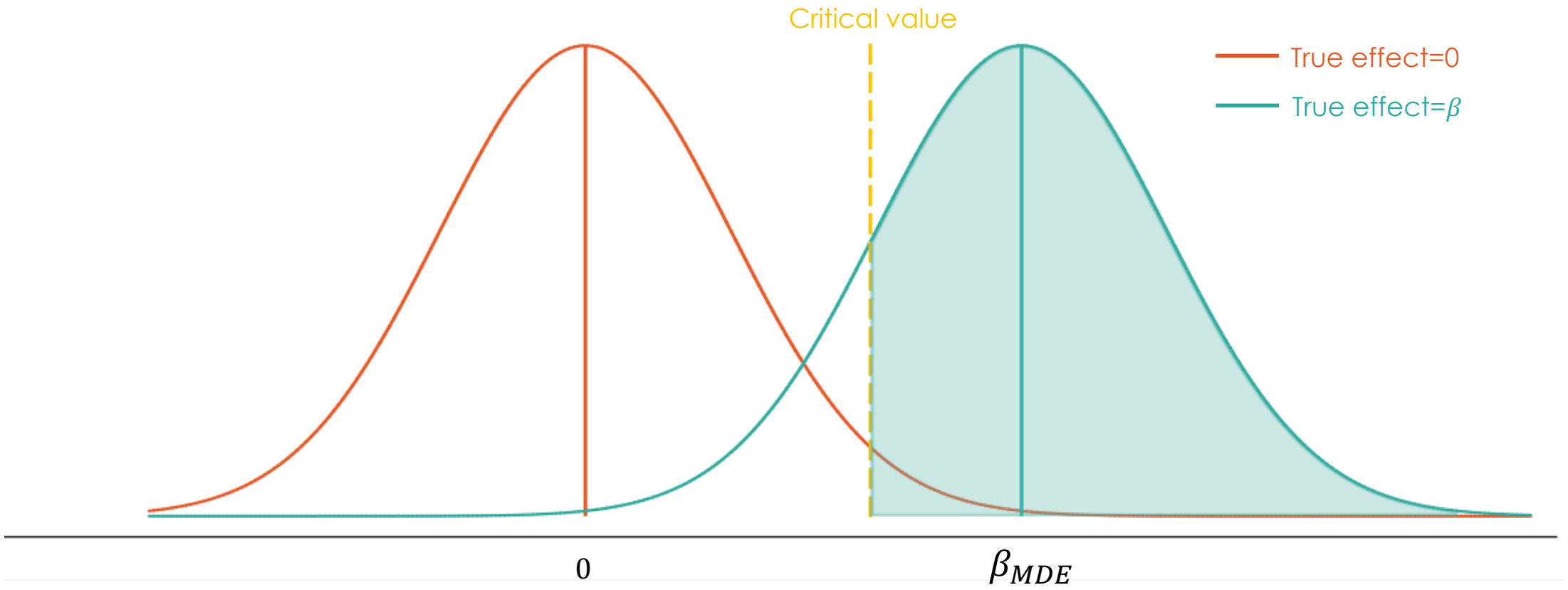
II. Hypothesis testing and statistical power

III. Power calculations

IV. Determinants of power

V. Practical tips

Minimum detectable effect size (MDE)



Minimum detectable effect (MDE)

Minimum Detectable Effect: The effect size that ensures that 80%* of the probability mass of the alternative distribution is to the right of the critical value.

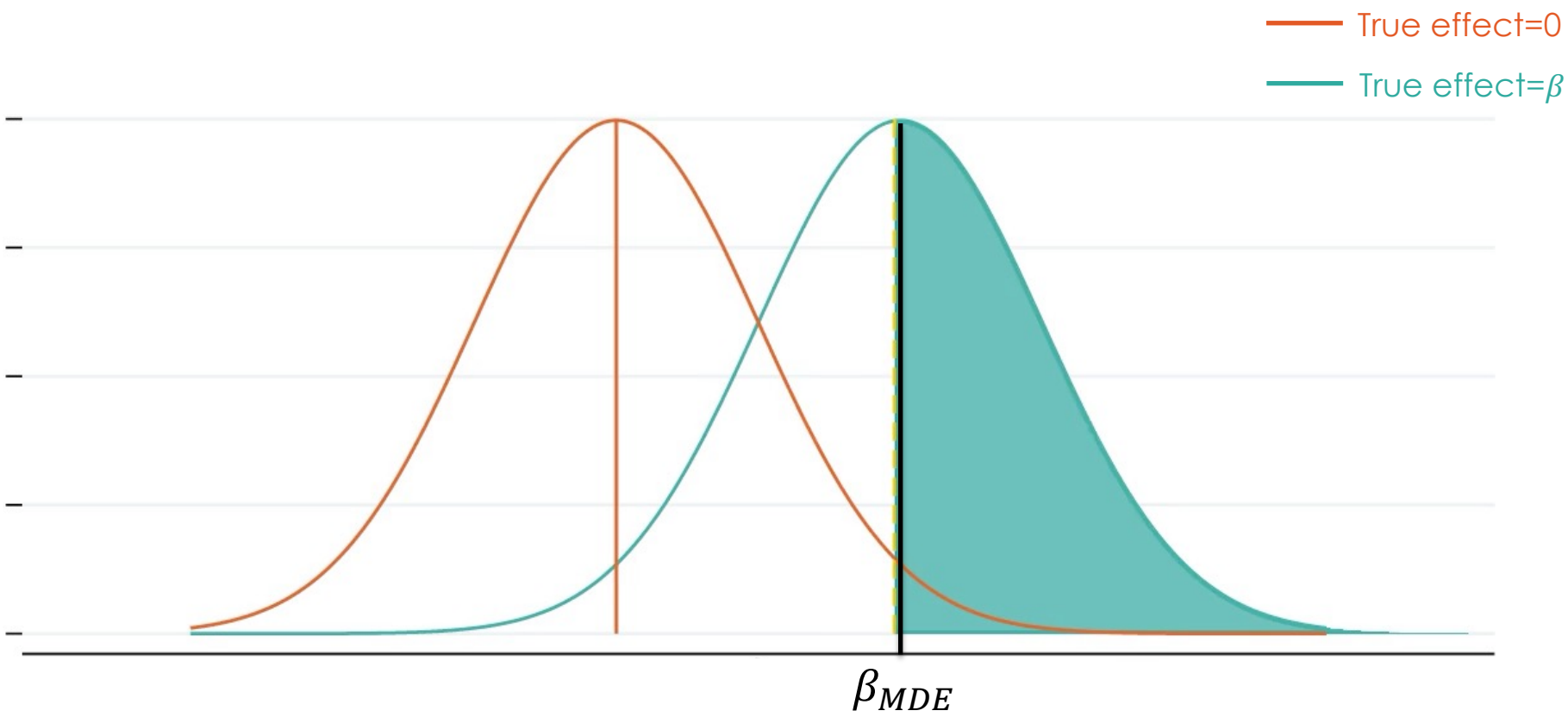
** Set by the researcher, so could also be 90% or other*

Power calculations: Two approaches

- **If sample size is flexible:** Calculate sample size that ensures 80% power for a given minimum detectable 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?

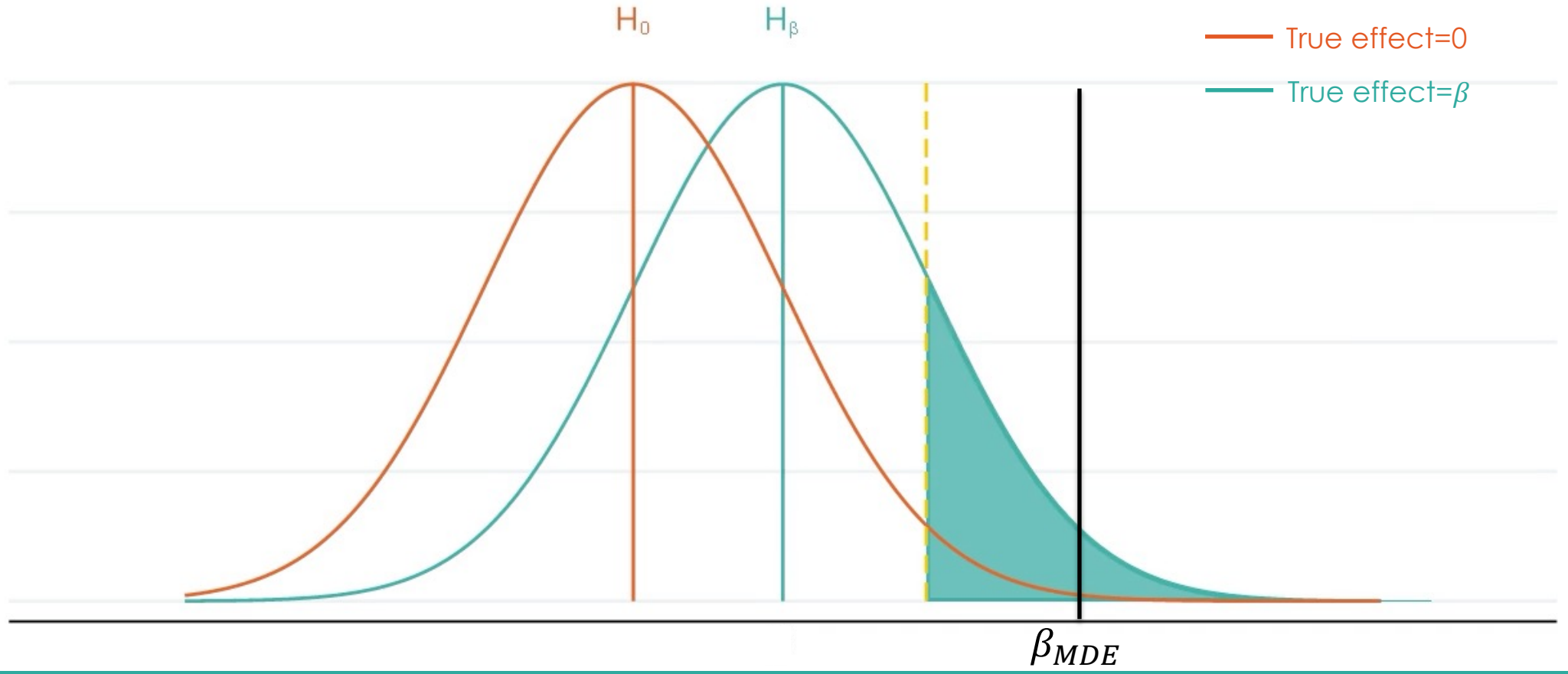
- **If sample size is fixed:** Calculate minimum true effect size required to achieve 80% power for a given sample size. Is this effect size reasonable?
  - What effects do similar studies find?
  - What effect would make the study cost-effective?
  - What effect would be required to be considered for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

Calculating required sample size for a given effect size



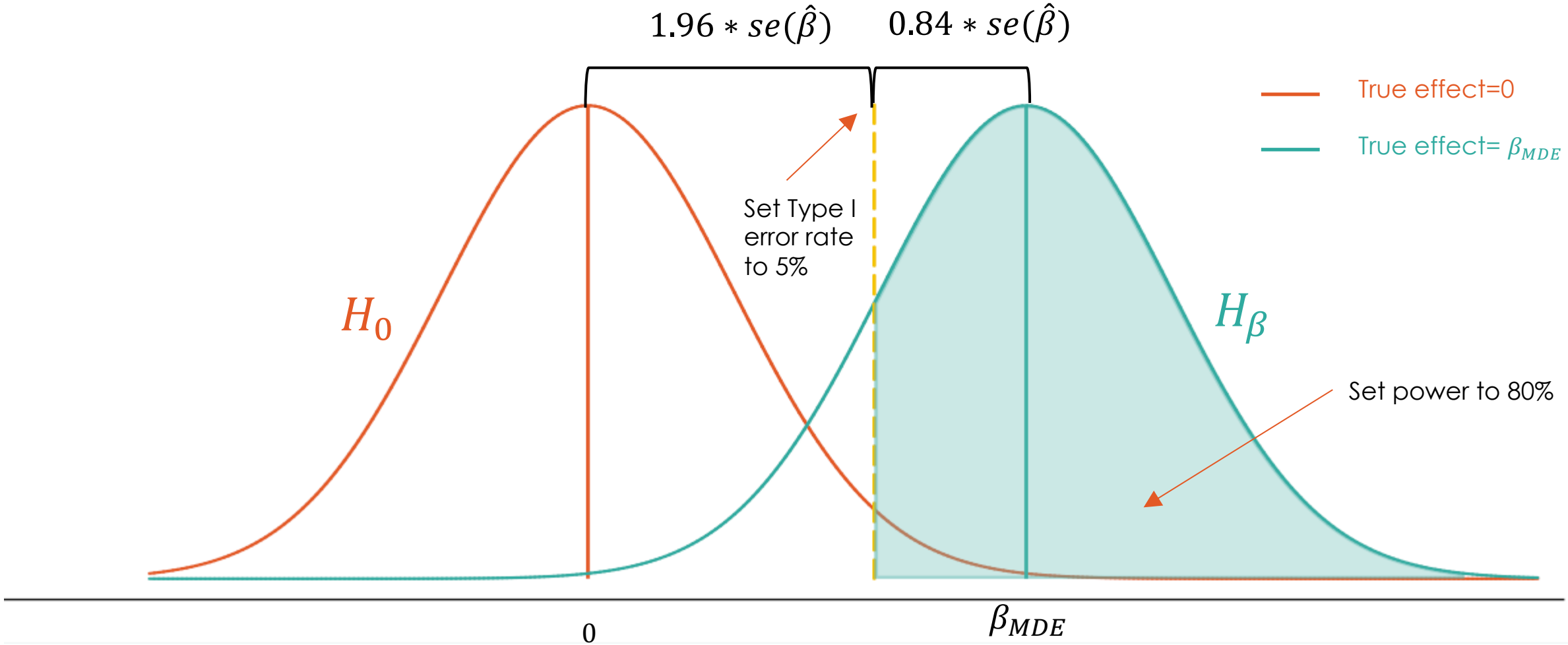
How narrow must the sampling distribution be for there to be 80% of the mass to the right of the critical value given the effect size?

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)



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

Calculating the minimum detectable effect size (MDE)

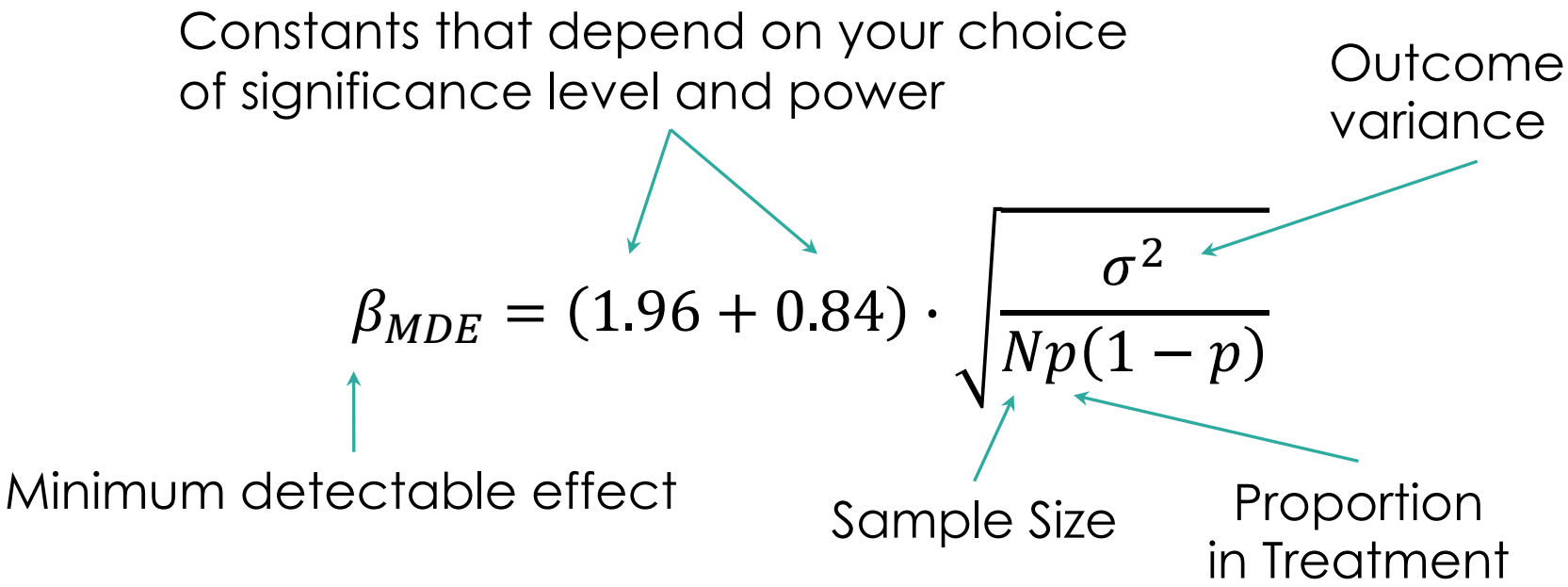
Constants that depend on your choice of significance level and power

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

Minimum detectable effect

Standard error of sampling distribution

Calculating the minimum detectable effect size (MDE)



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. Determinants of power

V. Practical tips

Power: Key determinants

- Minimum detectable effect: The minimum effect you will be able to detect
- Sample size: The number of units recruited to the study
- Pre-treatment outcome variance: How the outcome varies across units
- Sample split: The allocation across treatment and control groups
- The unit of observation and level of randomization

Power: Key determinants

- Minimum detectable effect: The minimum effect you will be able to detect
  - Compliance: The proportion of your sample who comply with their treatment allocation
- Sample size: The number of units recruited to the study
  - Attrition: The proportion of recruited units who end up falling out of your sample
- Pre-treatment outcome variance: How the outcome varies across units
- Sample split: The allocation across treatment and control groups
- The unit of observation and level of randomization

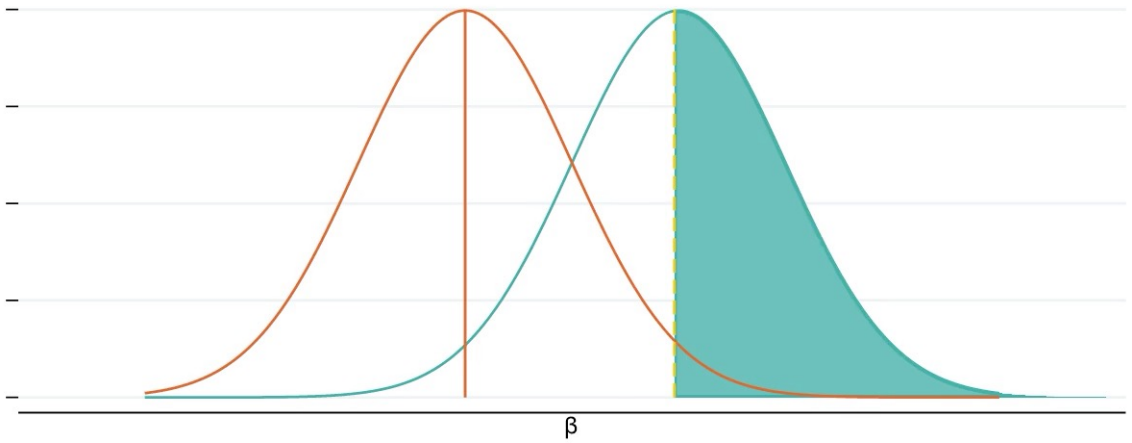
How does it affect power: Outcome variance

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

The pre-treatment outcome variance affects power through the shape of the sampling distribution

As the variance of the outcome decreases ↓, power increases ↑ because estimates become more precise

Why: It becomes easier to distinguish the part of the variation that comes from the program because there is less "noise" in the underlying data



How to determine: Variance can be estimated from baseline data, administrative data, or in datasets from similar studies

How does it affect power: Sample split

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

The sample split affects power through the shape of the sampling distribution

Power is maximized when the sample is split evenly between the treatment and comparison groups

Why: You want to maximize the sample size in both the treatment and comparison group at the same time

Important note: If total sample size can be increased by allocating more units to the comparison group, power might increase

- If intervention is expensive and data collection is cheap, consider allocating more units to comparison and increase *N*

How to determine: Chosen by the researcher based on feasibility and power considerations

How does it affect power: Compliance/take-up

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

The compliance/take-up rate affects power through the estimated effect size

As the compliance rate decreases ↓, power decreases ↓ because the estimated effect size decreases

Why: The estimated treatment effect size, $\hat{\beta} = \bar{Y}_T - \bar{Y}_C$, is unbiased only when comparing all in T to all in C regardless of whether they have been treated

- When not everyone is treated, $\bar{Y}_T \rightarrow \bar{Y}_C$

Important note: β_{MDE} increases at the rate of $\sqrt{N}$, so if compliance is 50%, you need 4x as many participants to retain the same power

- Increasing compliance is one of the strongest levers to increase power

How to determine: Knowledge of your sample, incentives, different studies

How does it affect power: Attrition

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

The attrition affects power through the effective sample size

As attrition rate increases ↑, power decreases ↓ because the effective sample size at endline decreases

Why: The estimated treatment effect size, $\hat{\beta} = \bar{Y}_T - \bar{Y}_C$, is based on endline values

- The effective sample depends on the number of people for whom there is endline data

Important note: If the attrition is correlated with the treatment allocation, the treatment effect estimate is no longer unbiased

- There has to be no differential attrition between the treatment and comparison groups

How to determine: Knowledge of your sample, incentives, different studies

How does it affect power: The unit of randomization

- 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)

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 unit of randomization affects power

- Samples with high intra-cluster correlation have similar individuals within clusters
  - Adding additional units from the same cluster adds less new information about the underlying distribution than adding a unit from a new cluster
  - Power increases ↑ as the number of clusters increase ↑
  - Power is relatively unaffected by the number of units within each cluster
  - 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 depends similarly on the number of clusters and units within clusters
  - ICC=0: You need as many units as you would need units if individually randomized

How to determine ICC: Estimate from baseline data, administrative data, or in available data from similar studies

Outline

I. Motivation

II. Hypothesis testing and statistical power

III. Power calculations

IV. Determinants of power

V. Practical tips

How can you improve the power of your study using the components we have discussed so far?

- Minimum detectable effect (incl. compliance)?
- Sample size (incl. attrition)?
- Pre-treatment outcome variance?
- Sample split?
- The level of randomization?
- Other design factors?

Tips for how to improve power I

- Increase sample size and take-up/compliance, reduce attrition, and conduct individual-level randomized studies when possible

- Add covariates (especially baseline measure of outcome of interest)
  - The variance included in the power calculated is the *residual variance* controlling for observable factors, so power increases as the explanatory power of covariates increases

- Reduce the number of treatment arms
  - The study needs to be powered for the smallest MDE among the intended treatment arm comparisons

Tips for how to improve power II

- Decrease the numbers of hypotheses you test (i.e., number of outcomes, number of subgroup analyses)
  - Study needs to be powered for the smallest MDE among the intended hypotheses
  - Plus, you need to adjust for multiple hypothesis testing

- Stratify the randomization on important observables
  - Randomizing within strata ensures baseline balance on important observables and (most likely*) increases power for subgroup analyses along these observables

*For more information, see Bruhn, Miriam and McKenzie, David, 2008. In Pursuit of Balance: Randomization is Practice in Development Field Experiments, Policy Research Working Paper #4752.

Tips for conducting power calculations

- Perform power calculations early – 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 how to allocate funds
- 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

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)

- 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)

Optimal design

- Optimal design and instructions
- Power calculations in Optimal Design (World Bank)

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., & 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..

Reuse and citation

To reference this lecture, please cite as:

J-PAL. "Lecture: Mechanics of Power." Abdul Latif Jameel Poverty Action Lab. 2023. Cambridge, MA

Appendix

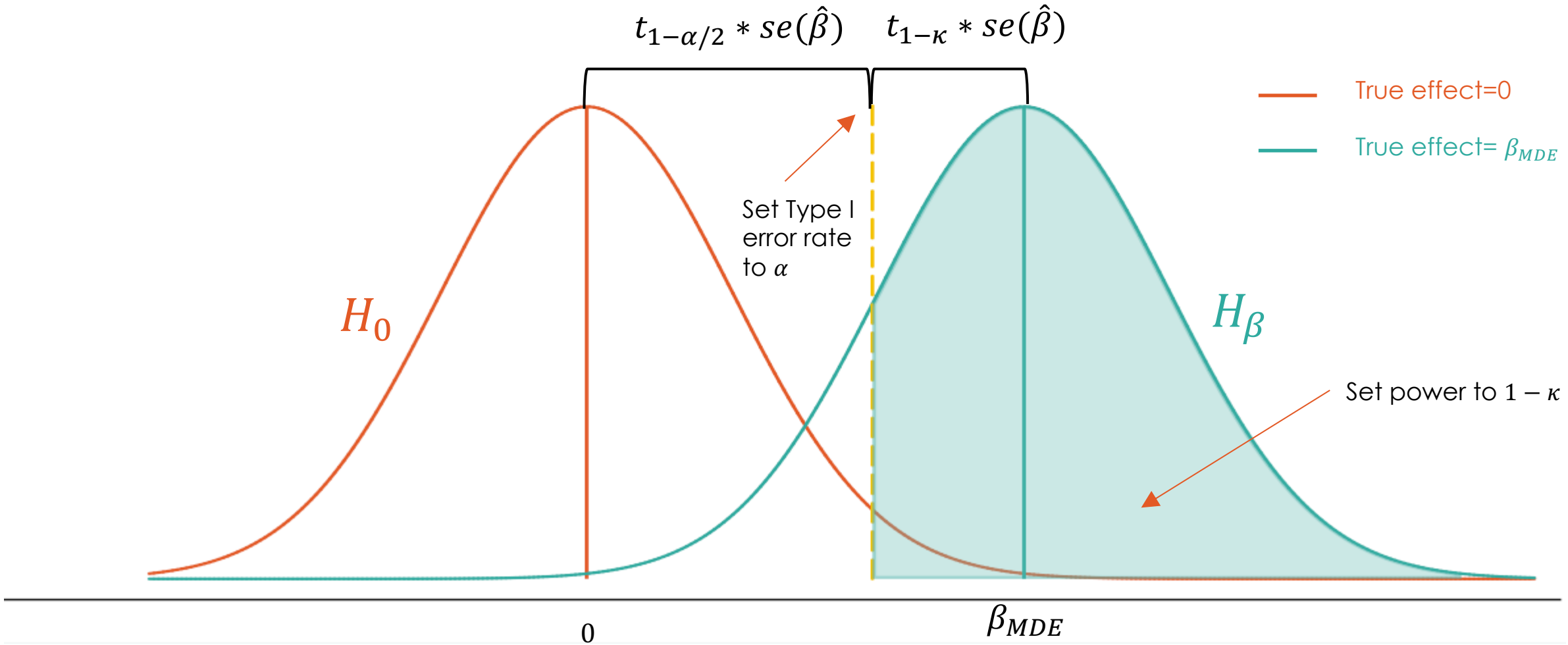
Power calculations step by step: Calculate MDE

1. Set desired power (e.g. 80%) and significance level (e.g. 5%)
2. Decide allocation ratio of the sample into treatment and control
3. Set sample size, number of clusters, and cluster size (if applicable) – this may be based on the budget and the design
4. Estimate variance & ICC (if applicable)
5. Back out the MDE for each outcome of interest, subgroup analysis, and comparison across treatment arms – adjust MDE based on expected compliance and attrition
6. Ask: Is the MDE realistic/policy-relevant

Power calculations step by step: Calculate sample size

1. Set desired power (e.g. 80%) and significance level (e.g. 5%)
2. Decide allocation ratio of the sample into treatment and control
3. Set MDE, adjusted by expected compliance and attrition rate
4. Estimate variance & ICC (if applicable)
5. Back out the sample size – if calculating number of clusters, specify cluster size, and vice versa
6. Conduct sensitivity analysis

Calculating minimum detectable effect (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

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

Minimal detectable effect

Sample Size

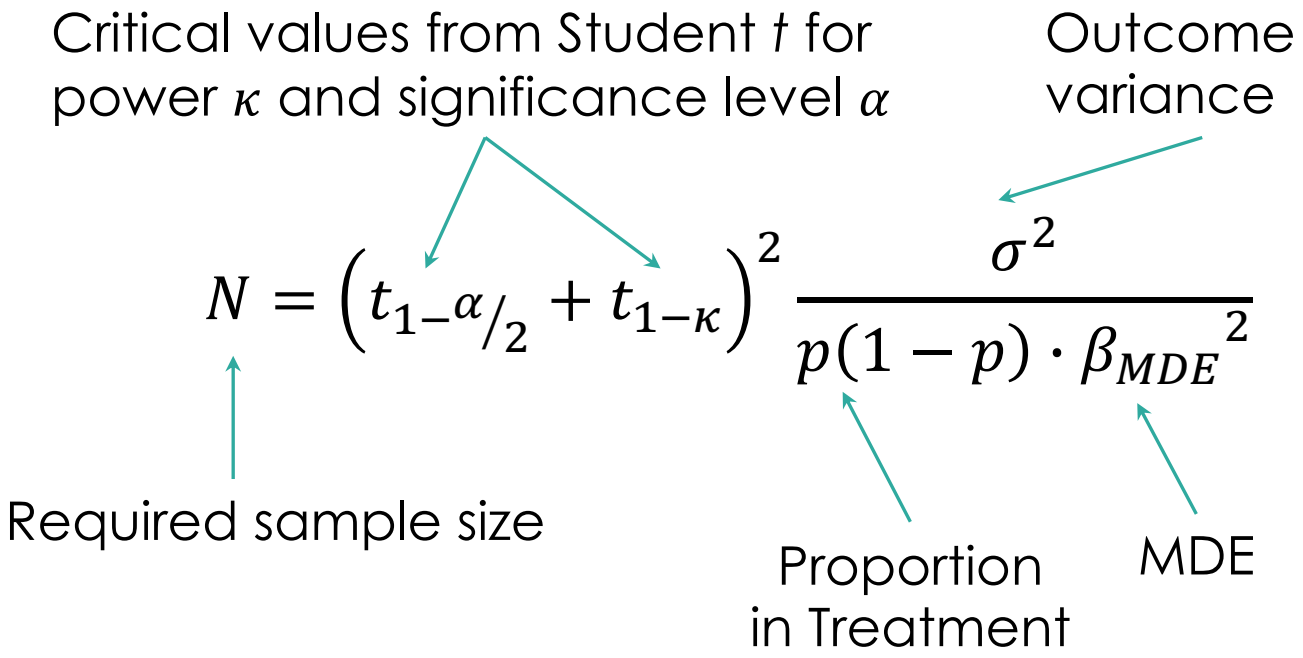
Proportion in Treatment

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



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*

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

Intra-cluster correlation coefficient

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

Minimal detectable effect

Number of clusters

Cluster size

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