J-PAL
ABDUL LATIF JAMEEL POVERTY ACTION LAB

Statistical Power and Choosing the Right Sample Size

J-PAL

Outline

I. True and Estimated Treatment Effect

II. Distribution of Treatment Effect Estimates

III. Hypothesis Testing

IV. Statistical Power and Sample Size

V. Power in Clustered Designs

VI. Calculating Power in Practice

Outline

I. **True and Estimated Treatment Effect**

II. Distribution of Treatment Effect Estimates

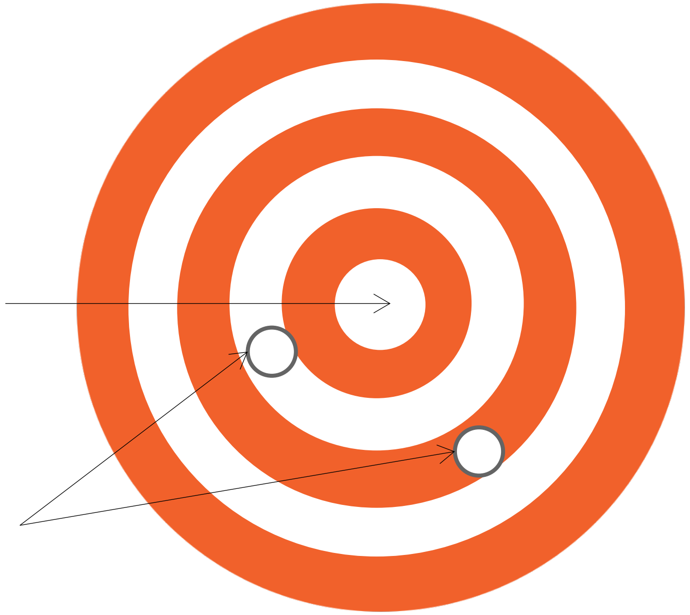
III. Hypothesis Testing

IV. Statistical Power and Sample Size

V. Power in Clustered Designs

VI. Calculating Power in Practice

Estimating the true treatment effect with an experimental sample.



The true treatment effect: difference in outcome with and without the program.

Estimates of the treatment effect from different samples

Average in treatment group

−

Average in control group

=

Estimated effect

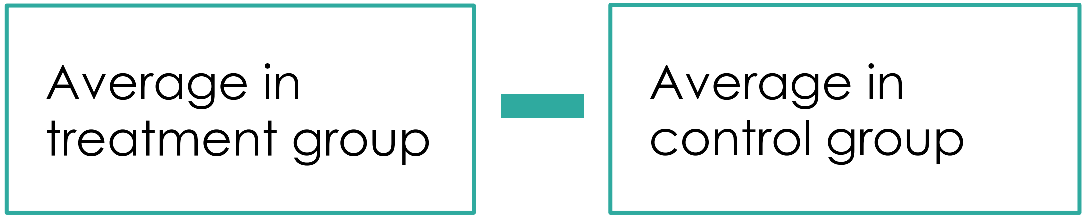
Chance variation from sampling implies that the estimated effect is not always exactly equal to the treatment effect. How do we address that?

Estimating the treatment effect from a sample

Estimate of the treatment effect

- The best estimator of the treatment effect is the difference in the average outcomes in a randomly selected treatment and comparison group.
- A treatment effect estimate is often denoted $\hat{\beta}$ (where β is the true effect).

Method 1: differencing



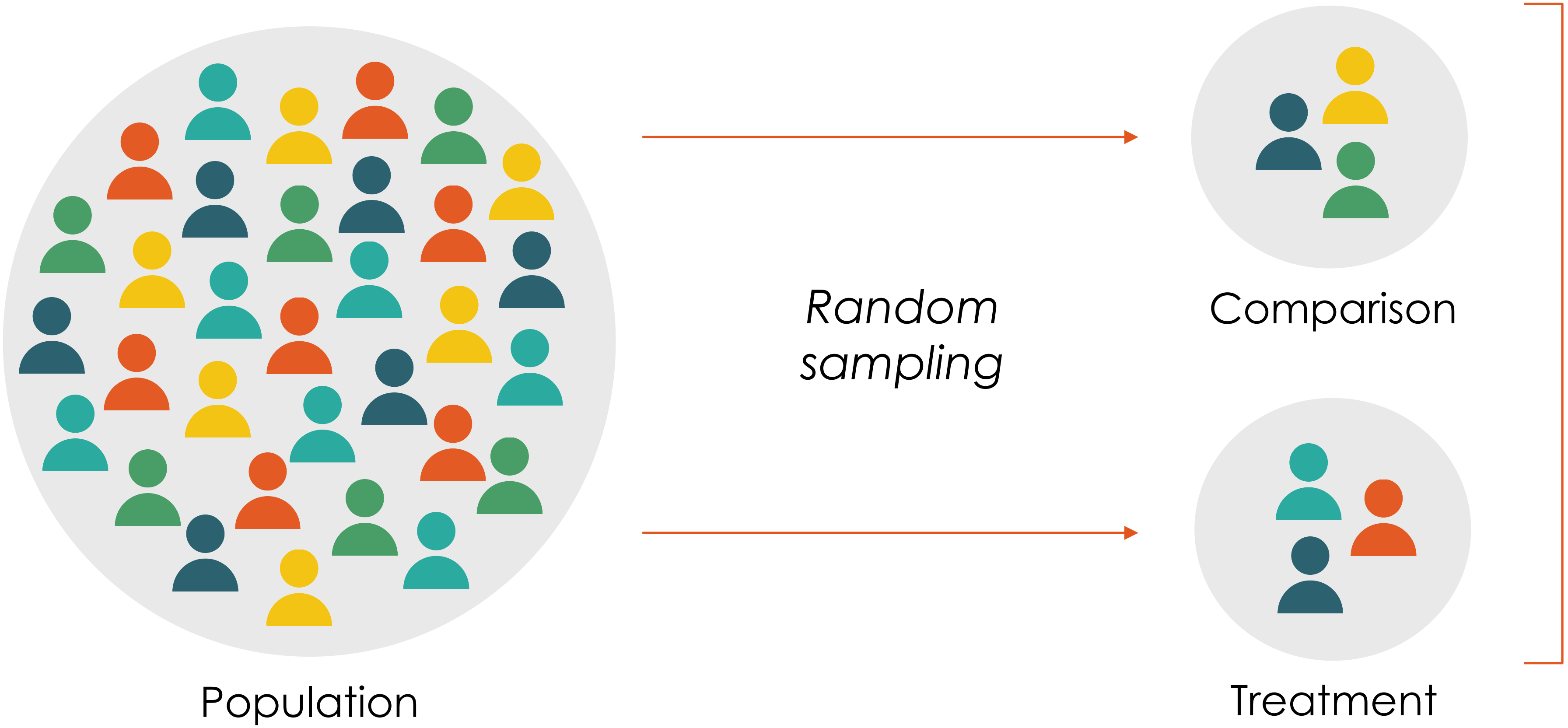
Average in treatment group − Average in control group

Method 2: regression

Use OLS to regress the outcome variable on the variable indicating treatment (equal to 1 if treated, equal to 0 if not treated)

Sampling variation and the treatment effect estimate

Randomization guarantees that treatment and control are not *systematically* different. They may still be different:



The treatment effect estimate will vary due to chance differences between the two groups.

Outline

I. True and Estimated Treatment Effect

II. Distribution of Treatment Effect Estimates

III. Hypothesis Testing

IV. Statistical Power and Sample Size

V. Power in Clustered Designs

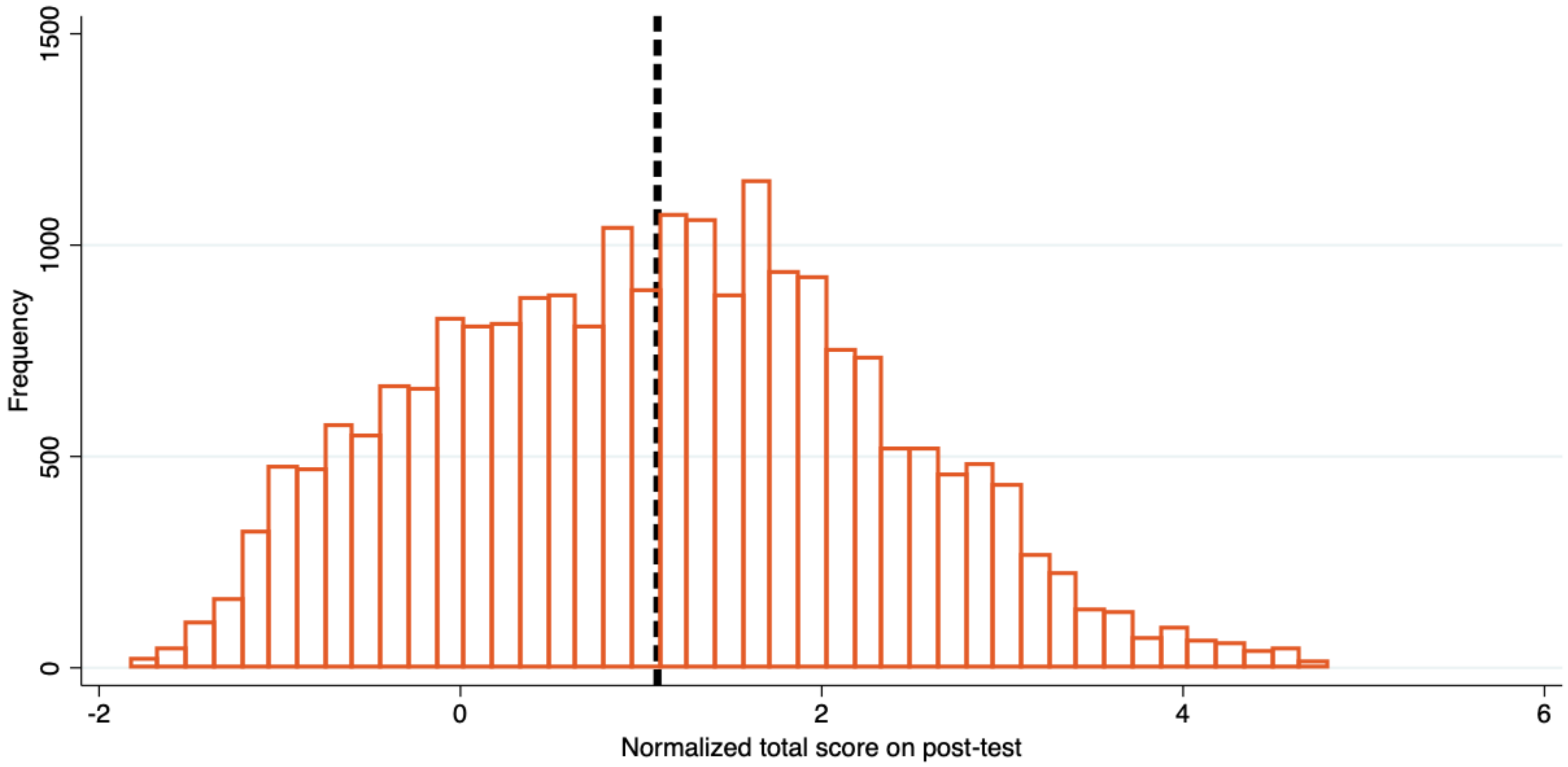
VI. Calculating Power in Practice

Example: evaluate the effect of a remedial education program in urban Indian schools on test scores.

Photo: Putul Gupta | J-PAL

Photo: Arvind Eyunni | Pratham

Example: the distribution of standardized school test scores in urban Indian schools.

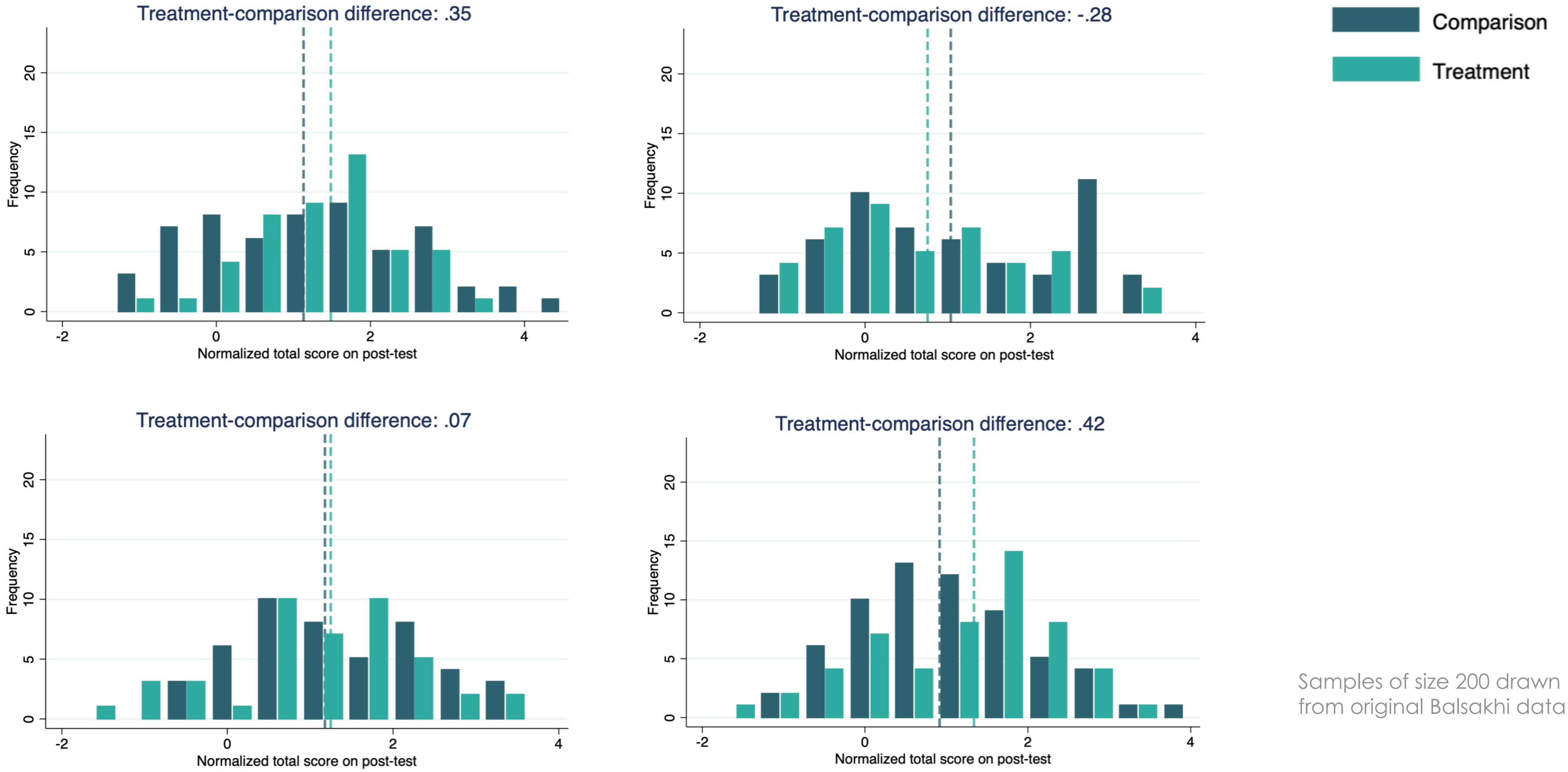


Source: Banerjee, Abhijit; Cole, Shawn; Duflo, Esther; Linden, Leigh, 2017, "Balsakhi", Harvard Dataverse,

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

A sample of over 23000 students shows the natural variation in standardized test scores (data from the "Balsakhi" remedial education program).

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



Treatment-comparison difference: .35

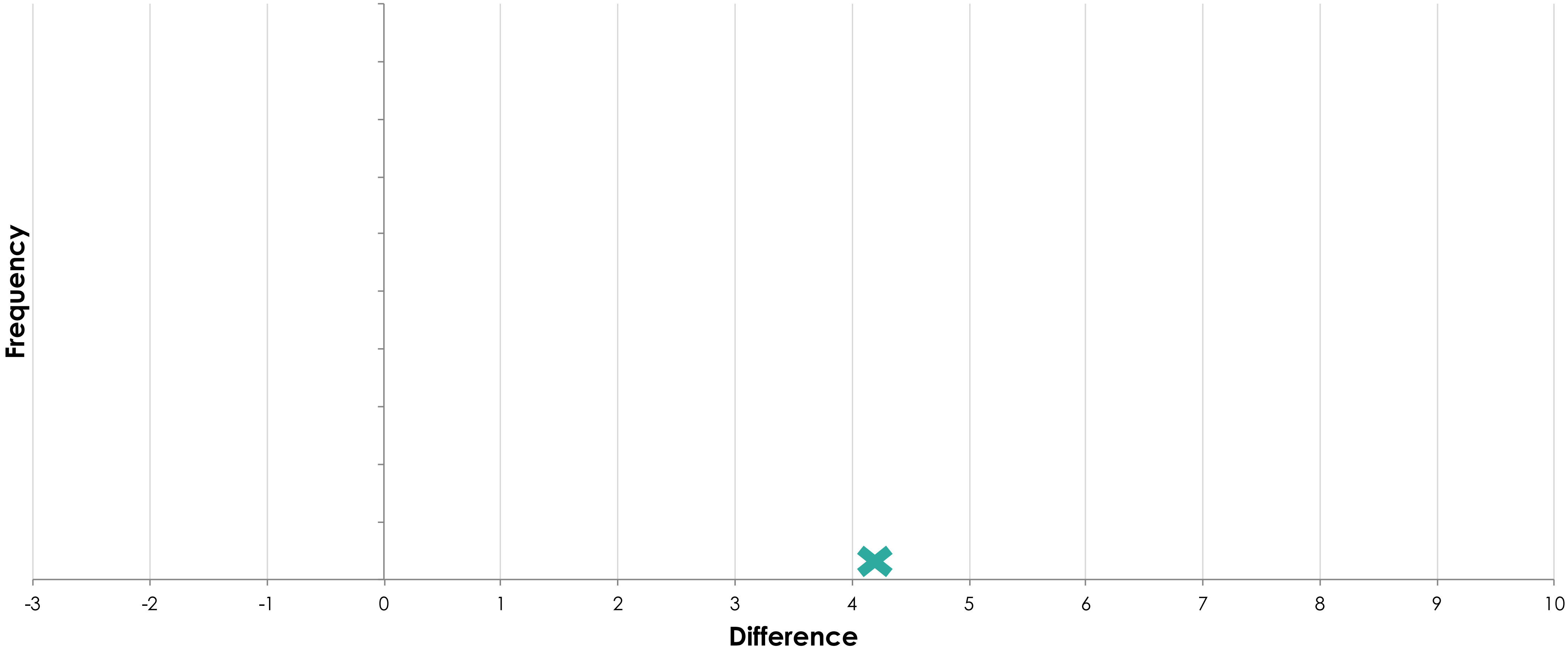
Treatment-comparison difference: -.28

Treatment-comparison difference: .07

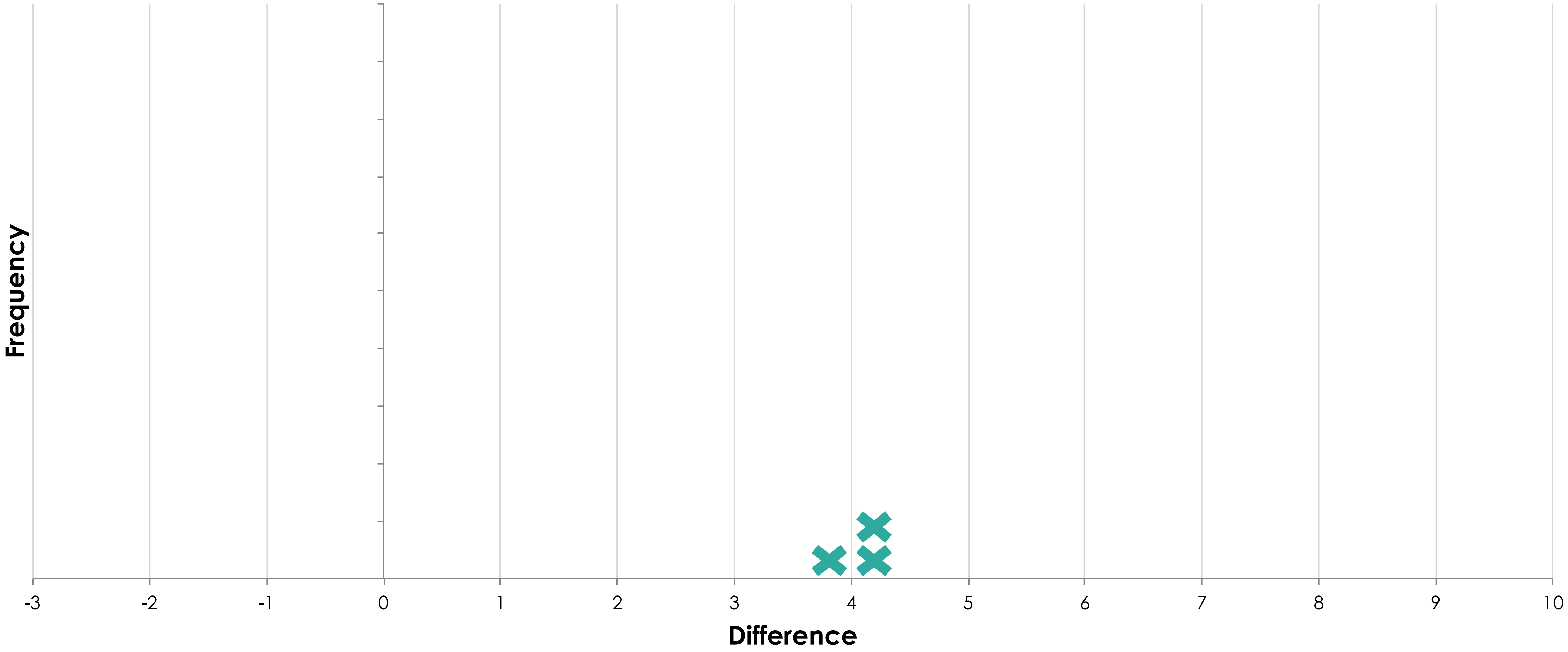
Treatment-comparison difference: .42

Samples of size 200 drawn from original Balsakhi data.

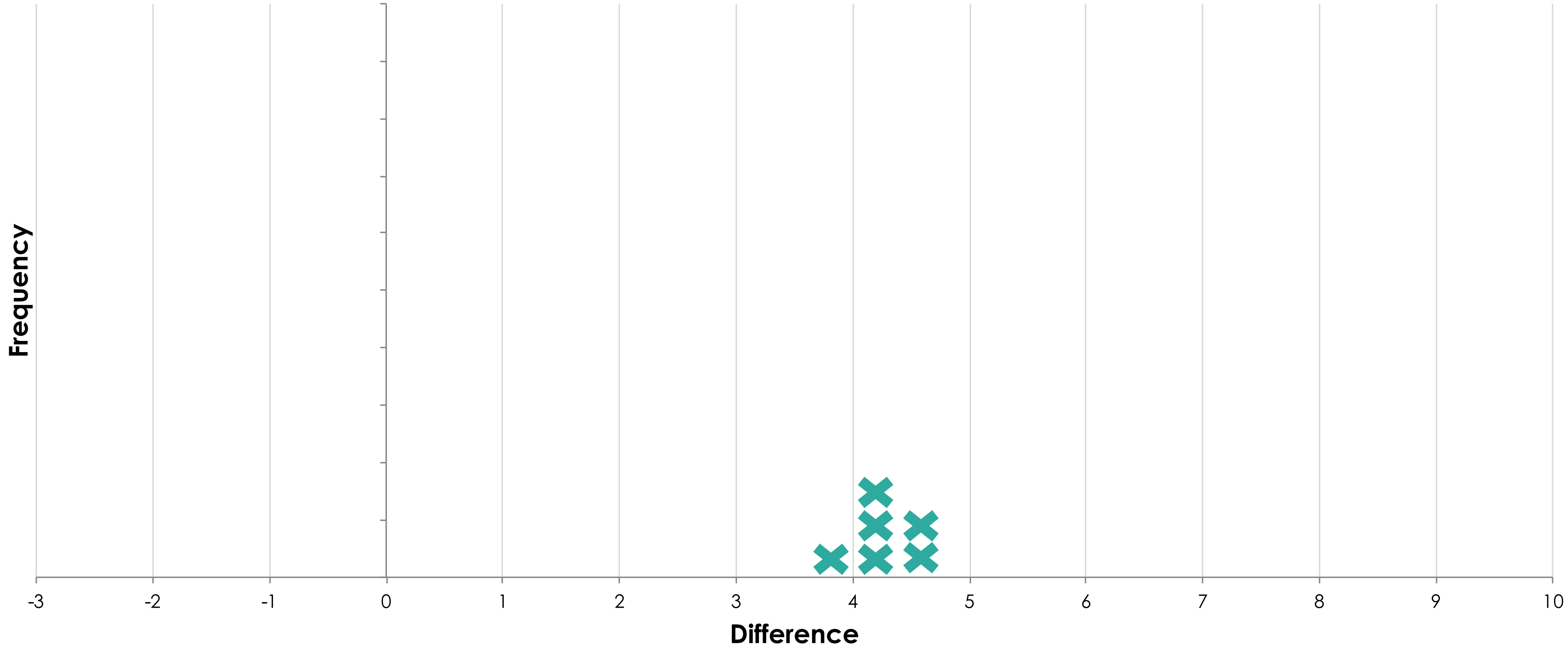
Estimated impact from one experiment



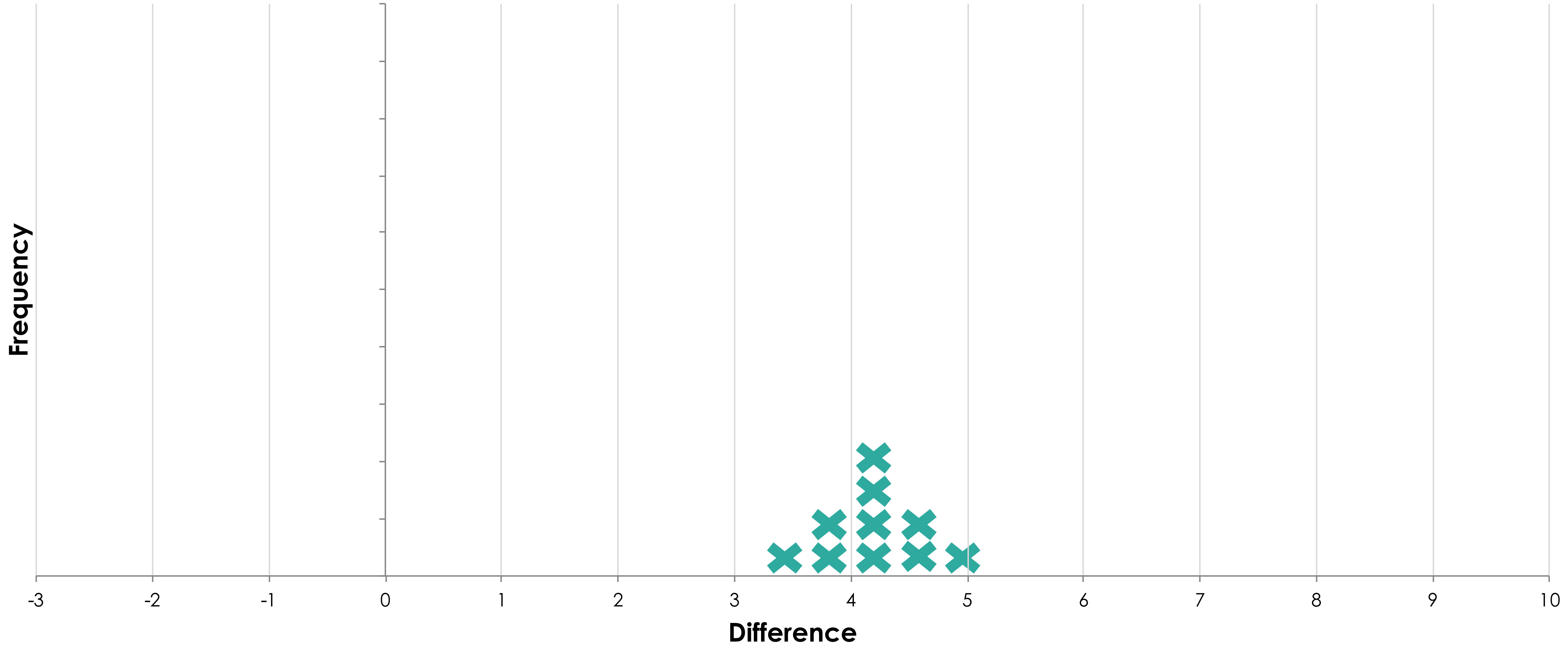
Three experiments, three estimated impacts



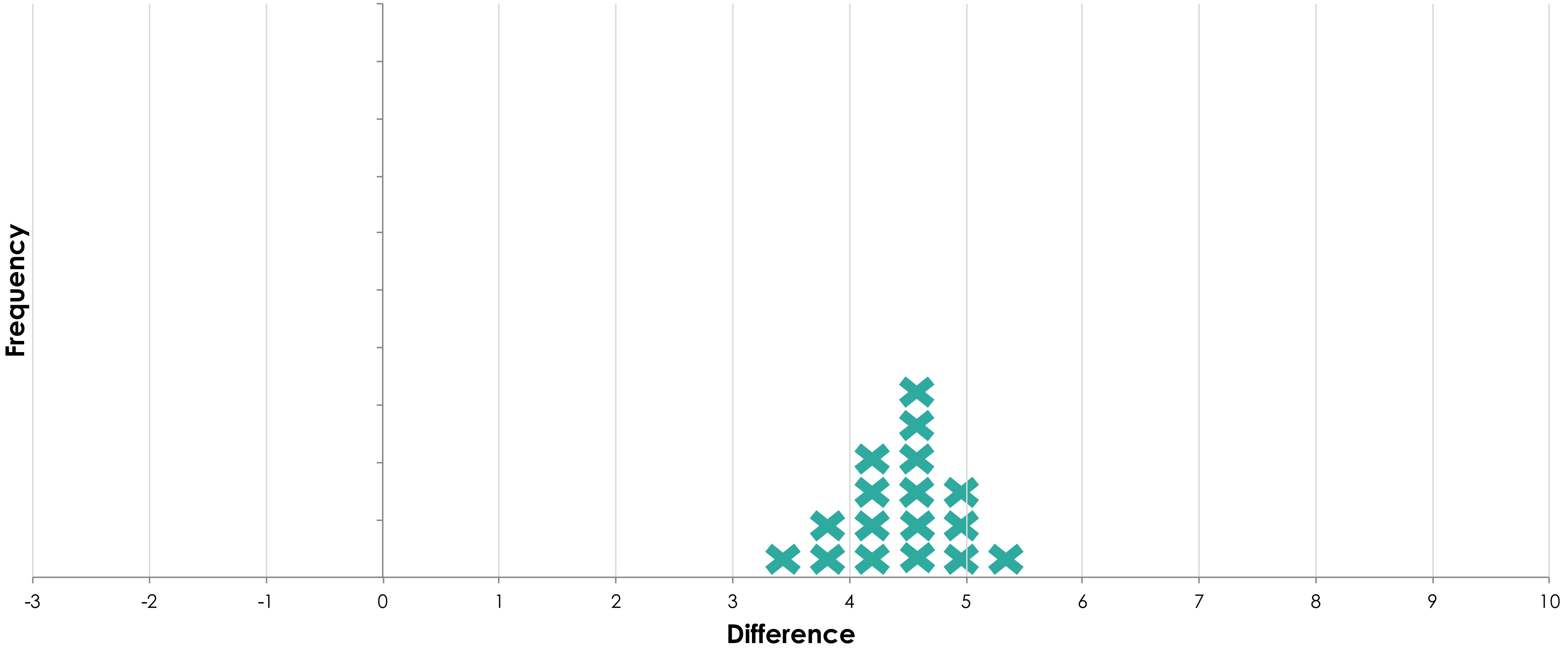
Six experiments, six estimated impacts



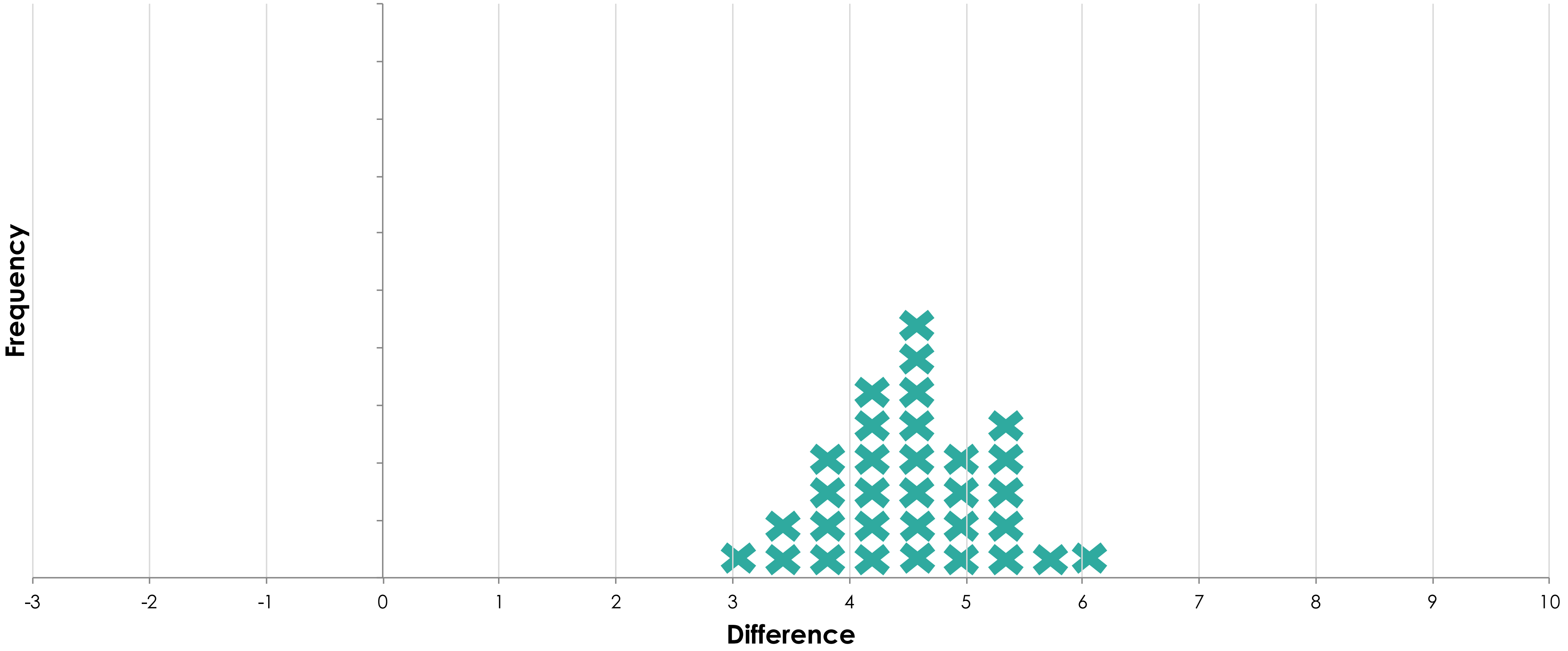
Many experiments: a distribution of estimates



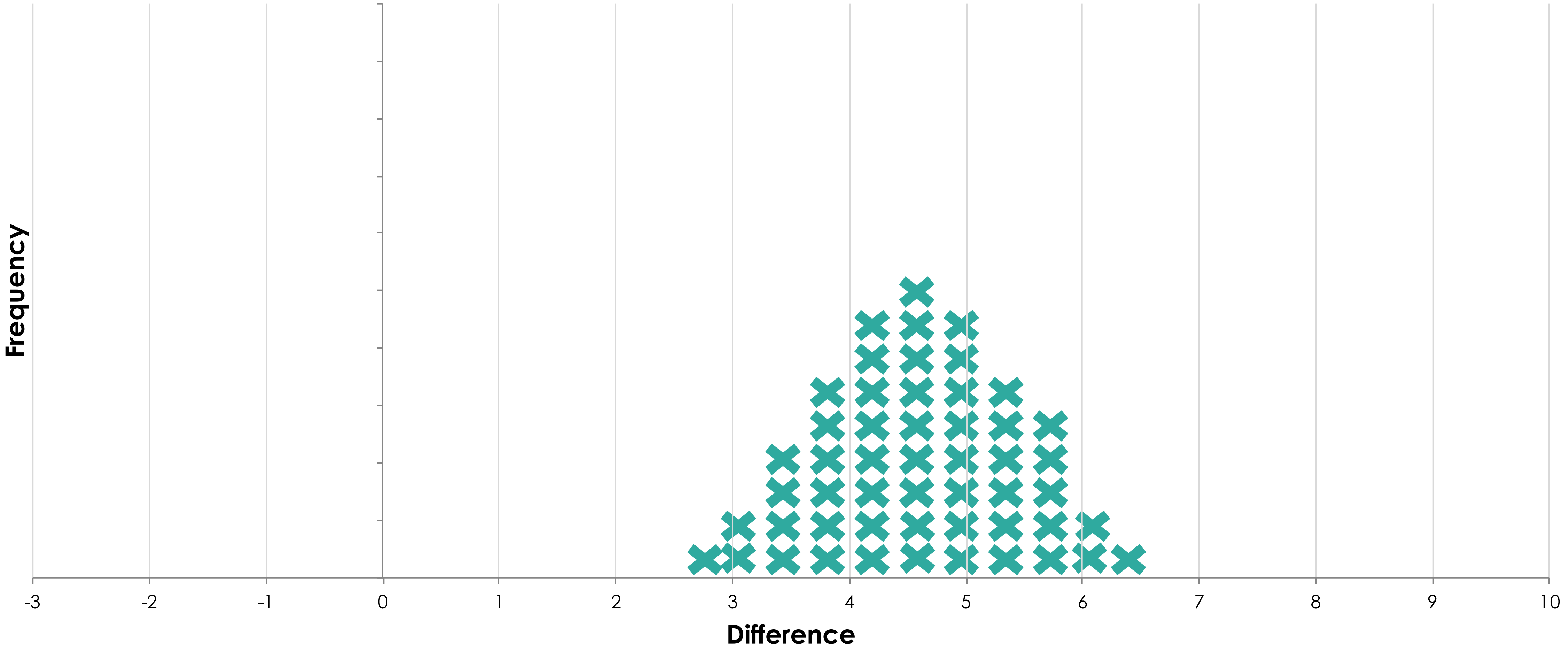
Many experiments: a distribution of estimates



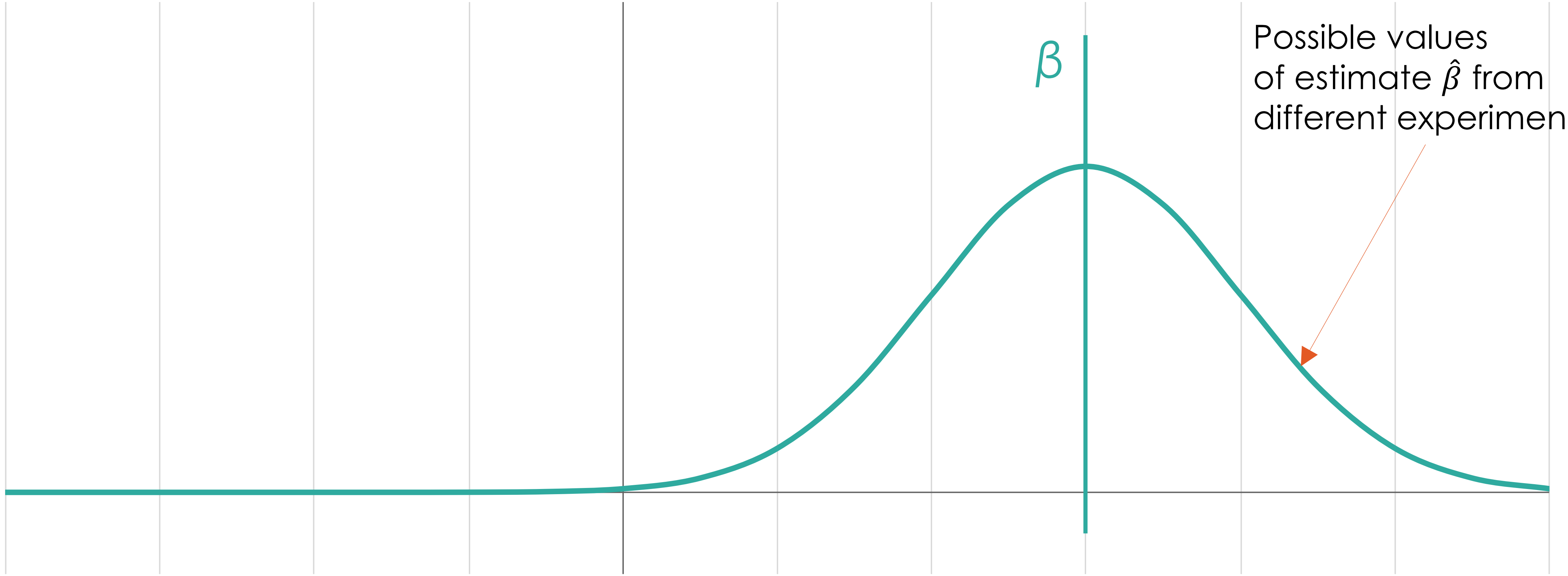
Many experiments: a distribution of estimates



Many experiments: a distribution of estimates



Distribution of estimates if the true effect is β

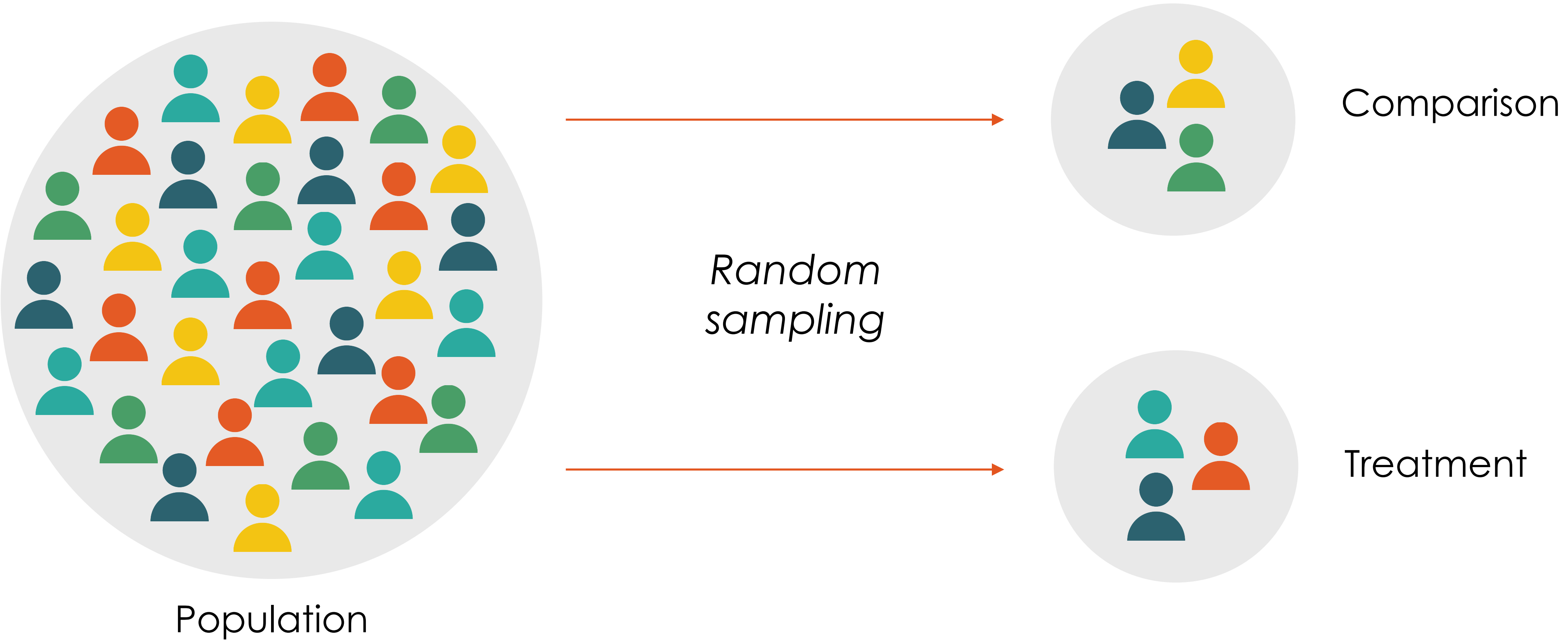


Possible values of estimate $\hat{\beta}$ from different experiments

Randomization guarantees that the estimated effects are normally distributed around effect β.

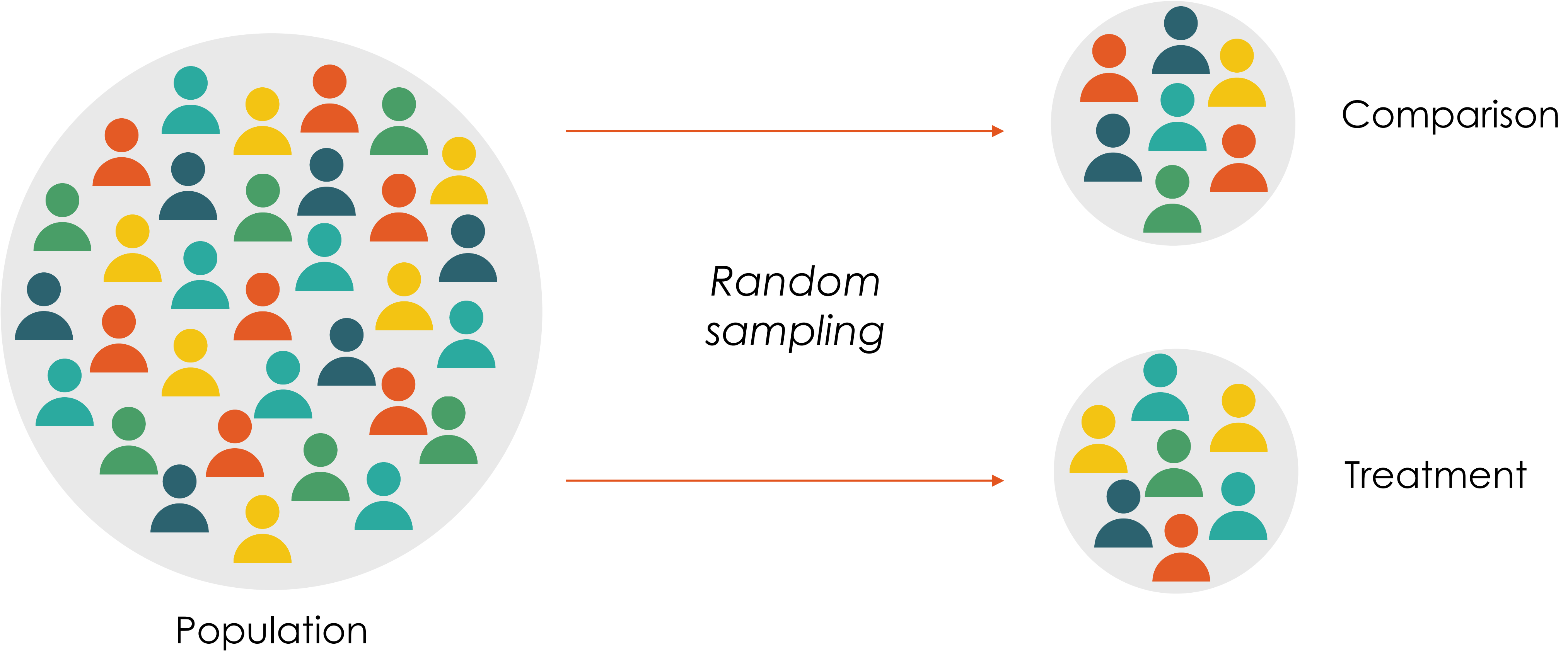
Intuition: how does sample size affect the treatment effect distribution?

Any two small samples may be quite different from each other.

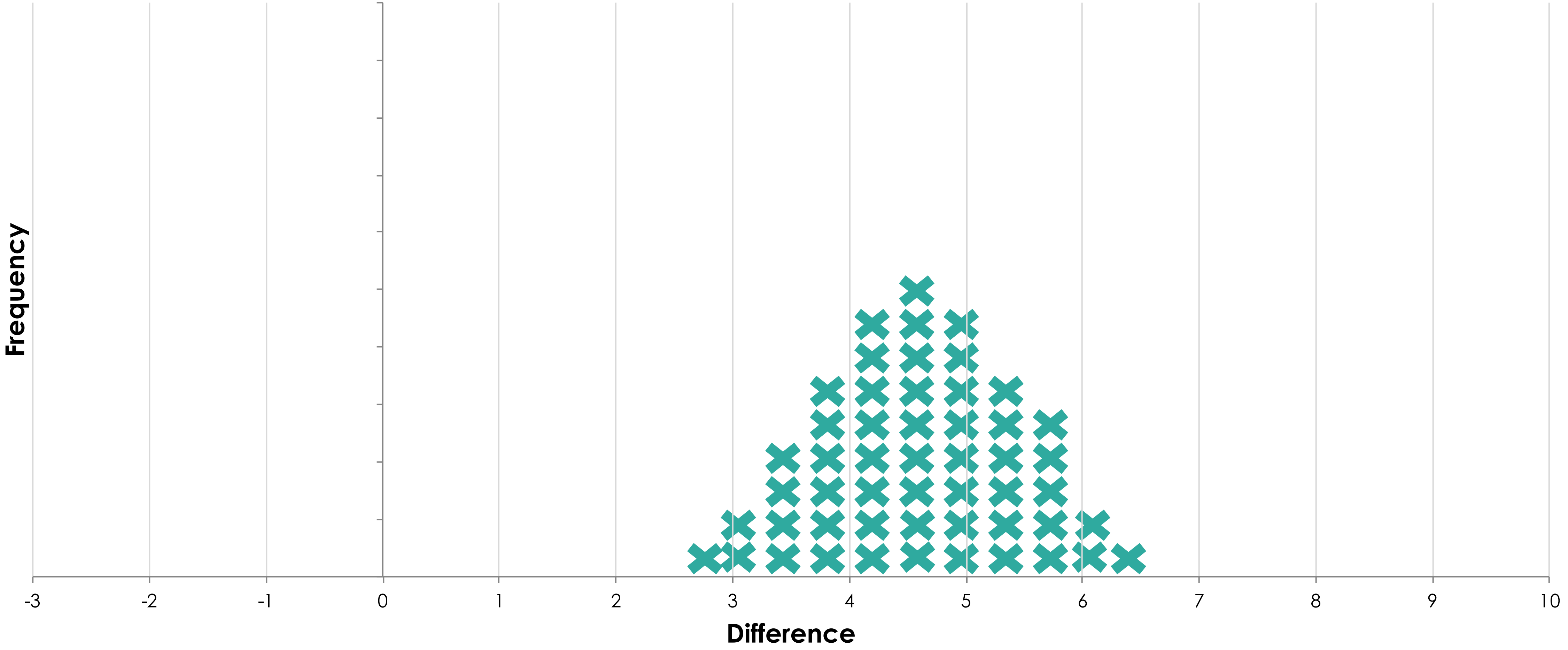


Intuition: how does sample size affect the treatment effect distribution?

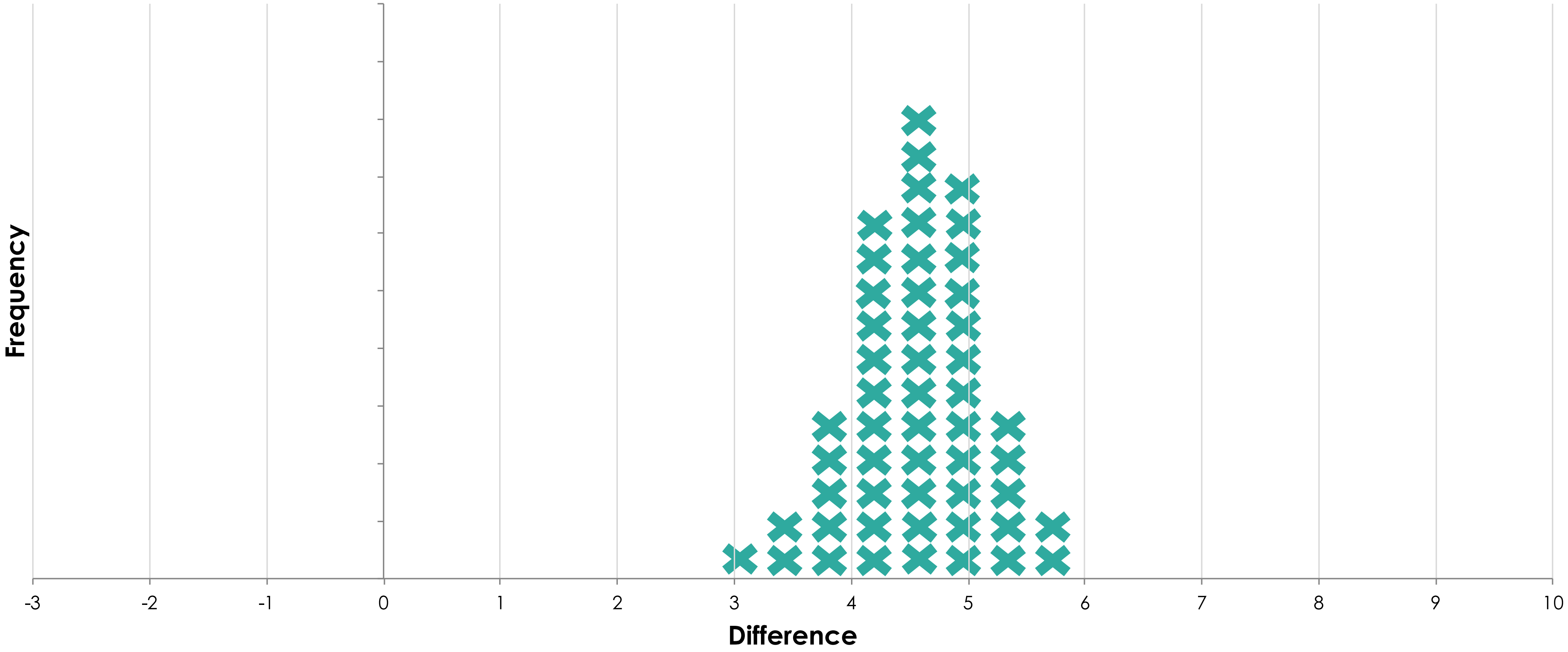
Larger samples tend to be more representative of the population and more similar to each other.



If treatment and control group are more representative, effect estimates from different samples vary less.

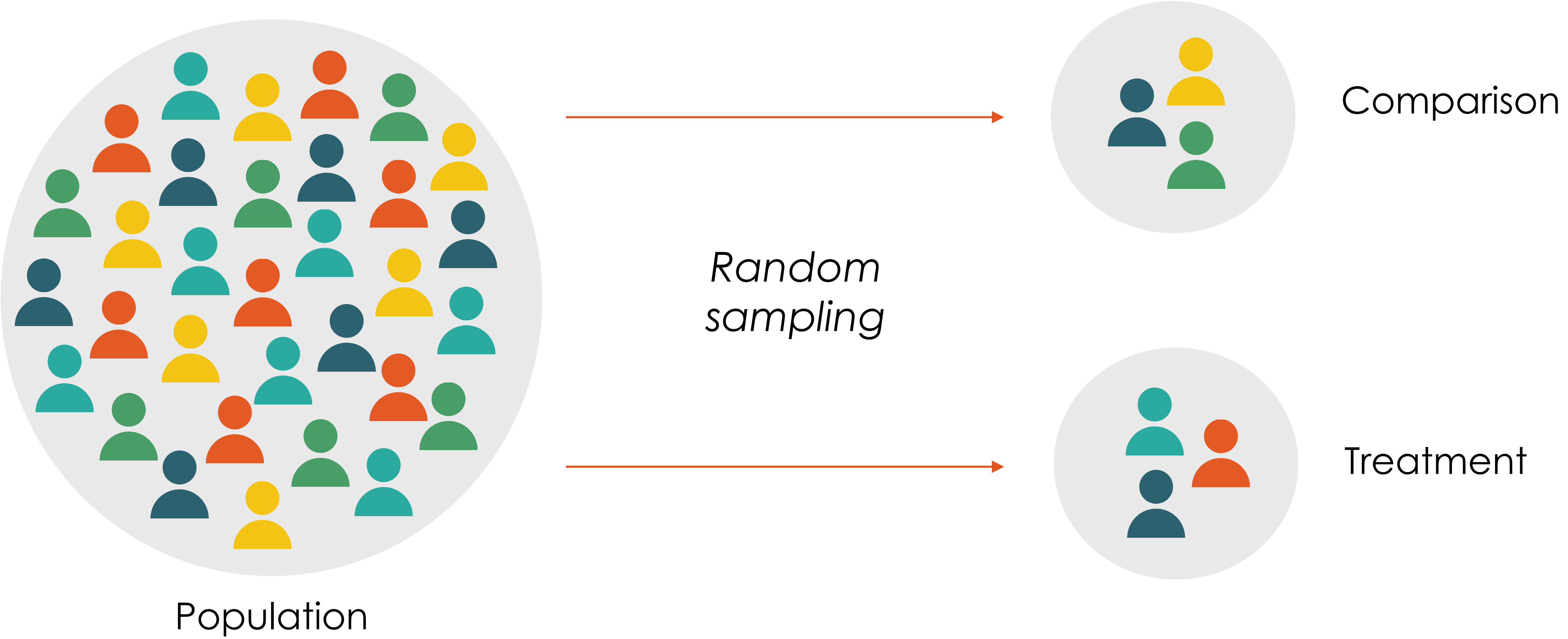


If treatment and control group are more representative, effect estimates from different samples vary less.



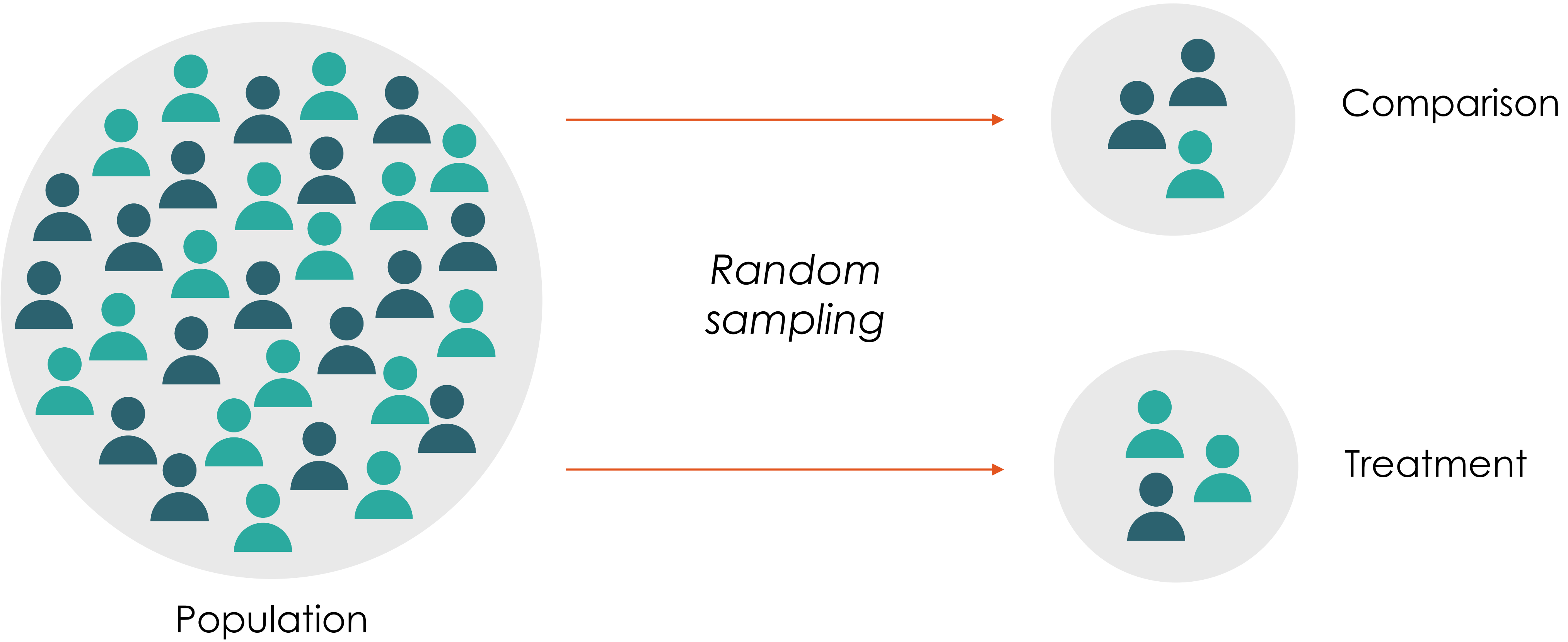
Intuition: how does variation in the underlying population affect the treatment effect distribution?

Samples from a population with high outcome variation can look quite different.



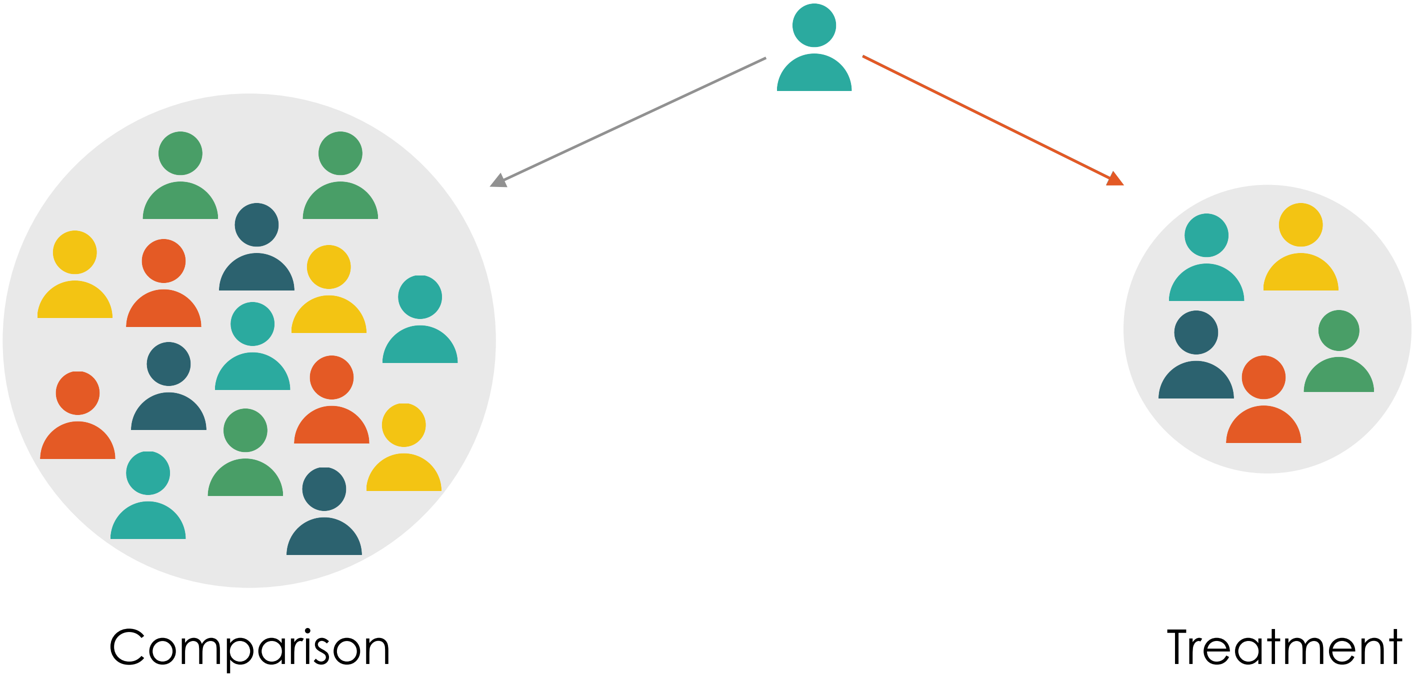
Intuition: how does variation in the underlying population affect the treatment effect distribution?

Samples from a population with little outcome variation tend to be more similar to the population as a whole and to each other.

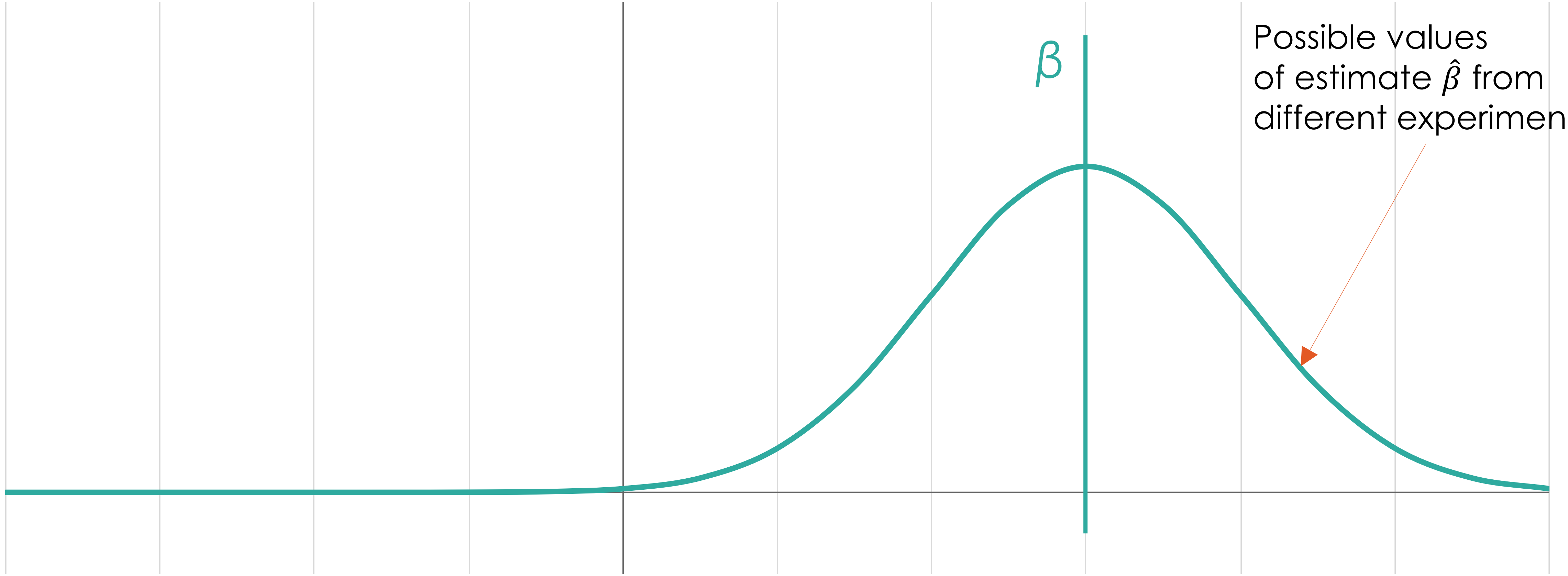


Intuition: how does the sample split into treatment and control affect the treatment effect distribution?

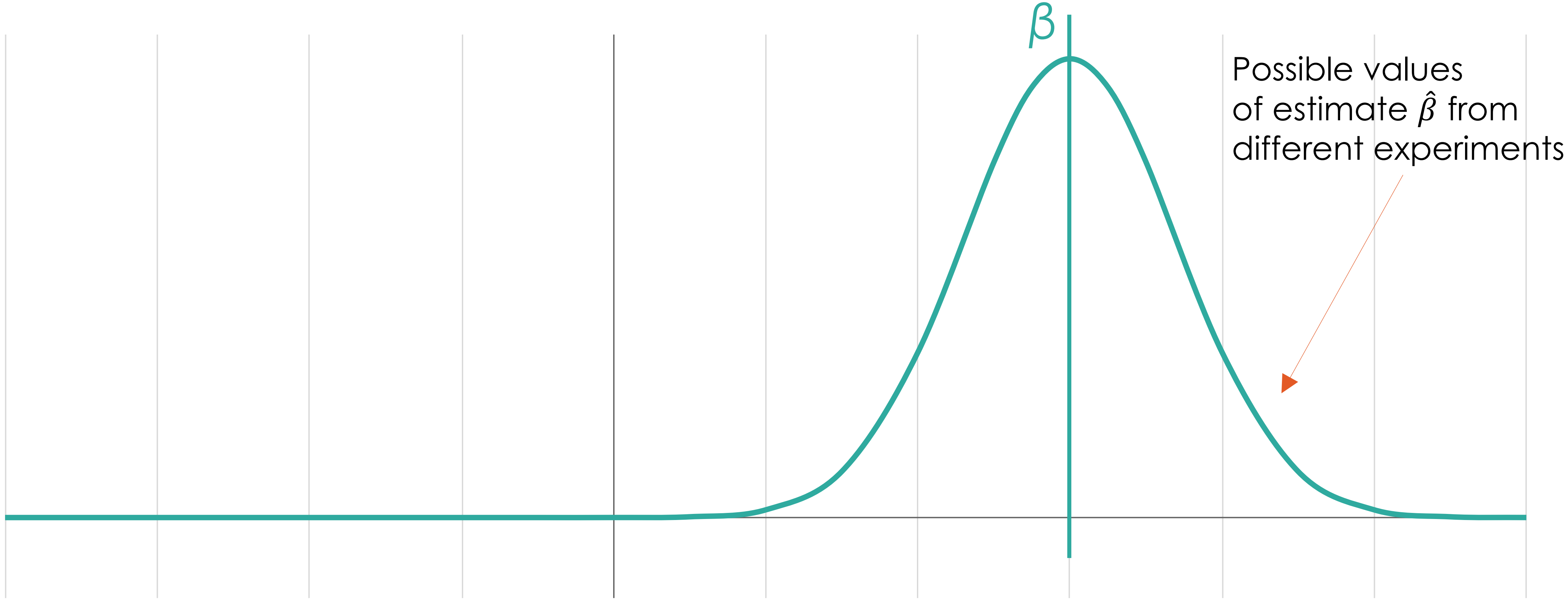
An additional subject in the smaller group – a more even split – reduces the sampling variation by more.



If treatment and control group are more representative, effect estimates from different samples vary less.



If treatment and control group are more representative, effect estimates from different samples vary less.



Larger samples, lower outcome variation, or a more even sample split: estimate distribution is more concentrated around true β.

Outline

I. True and Estimated Treatment Effect

II. Distribution of Treatment Effect Estimates

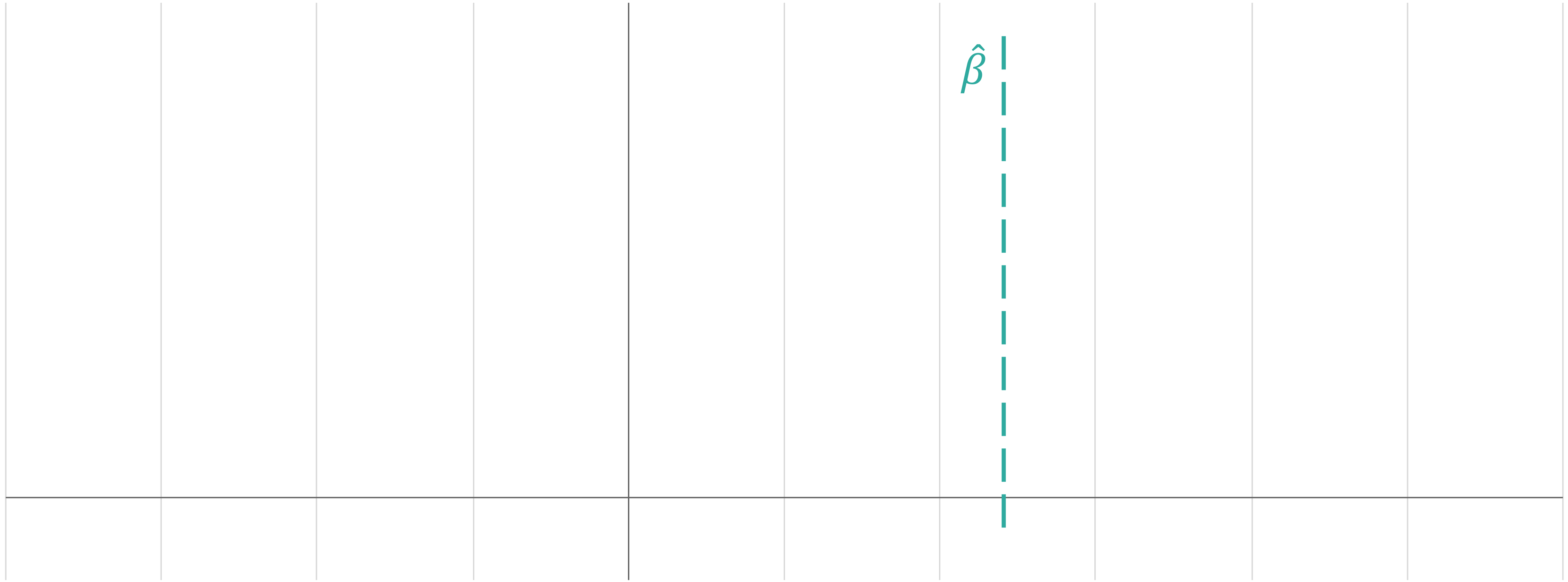
III. Hypothesis Testing

IV. Statistical Power and Sample Size

V. Power in Clustered Designs

VI. Calculating Power in Practice

Problem: we only see one estimate, from the experiment at hand. What do we conclude?



The true treatment effect β is unknown.

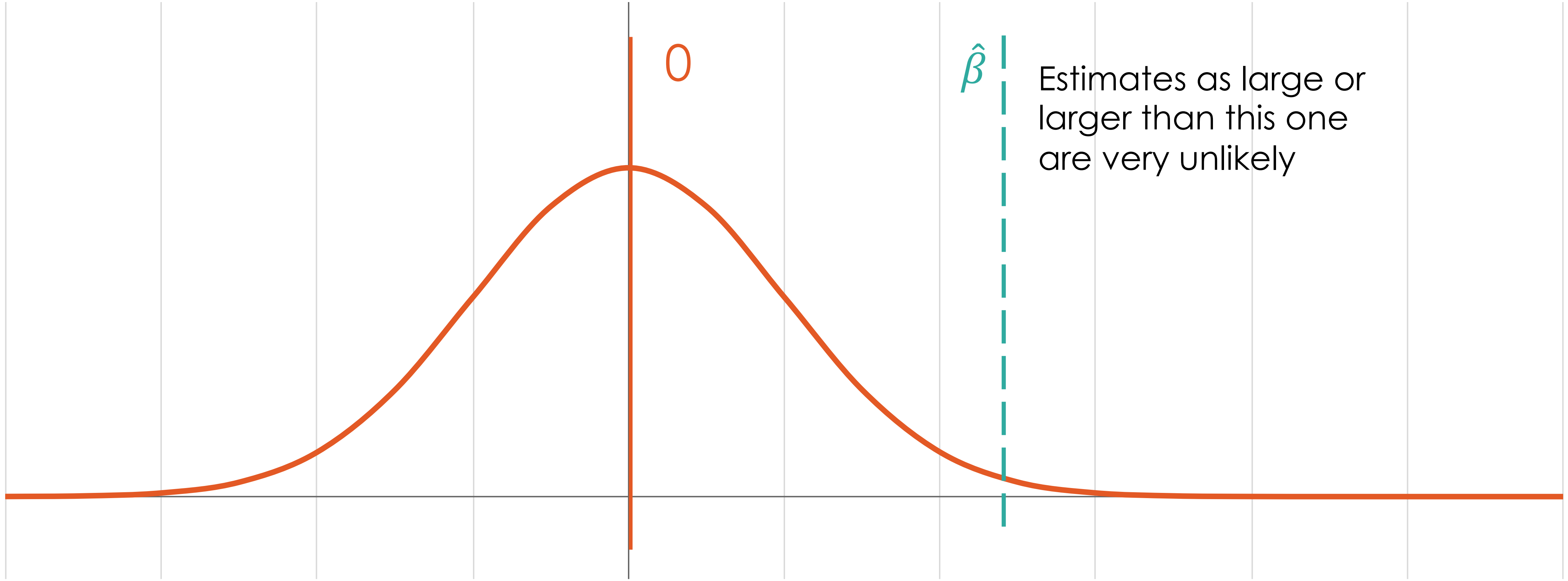
Hypothesis testing

Null hypothesis

The null hypothesis (H_0) is that there was no (zero) impact of the program on the outcome variable, i.e. $\beta = 0$.

- 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 $\widehat{\boldsymbol{\beta}}$ in an experiment, if the true effect was actually zero?**
- If it is "very unlikely", we can reject the null hypothesis.

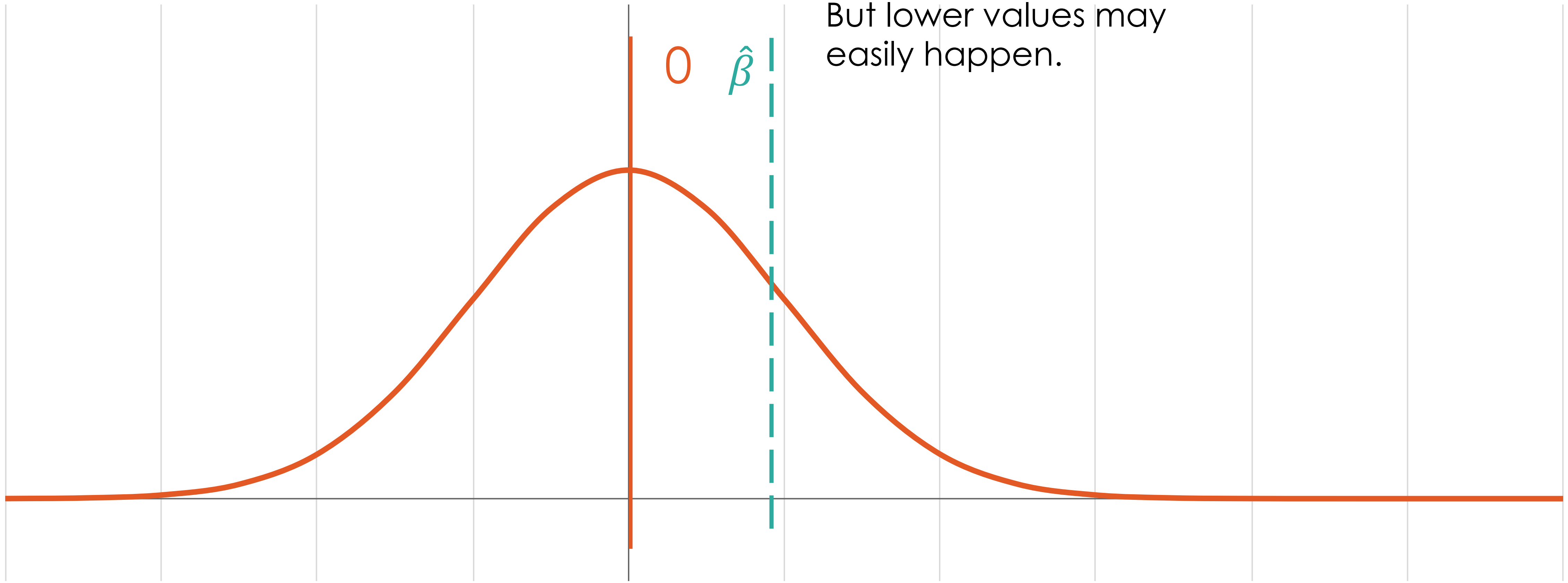
Distribution of estimates if the true effect is zero



Estimates as large or larger than this one are very unlikely

Assume a treatment effect of $\beta = 0$.

Distribution of estimates if the true effect is zero



But lower values may easily happen.

Assume a treatment effect of $\beta = 0$.

Definition: Significance Level

Type I error (false positive)

The probability of falsely concluding that there is a treatment effect.

The probability of rejecting the null hypothesis β = 0, even though the null hypothesis is true.

With any estimate, there will be some chance of a Type I error.

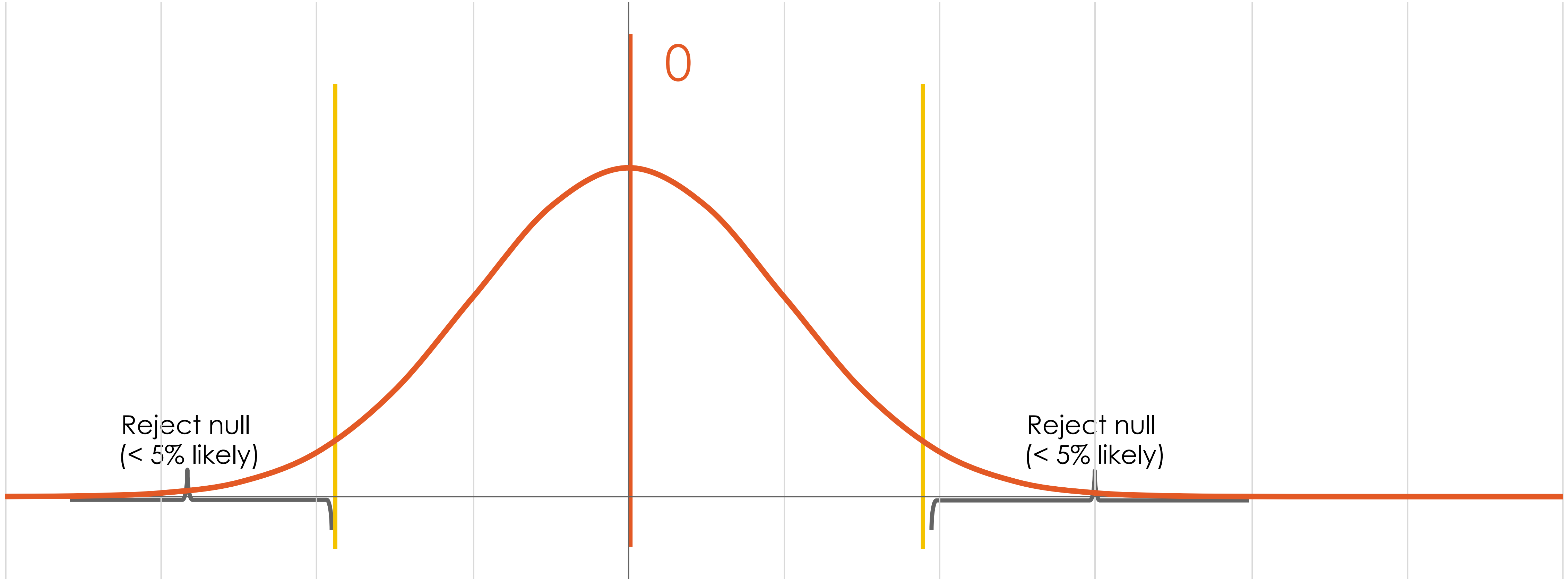
Significance level

The maximal probability of a Type I error we want to allow.

5% is most commonly used, but also 1% and 10%.

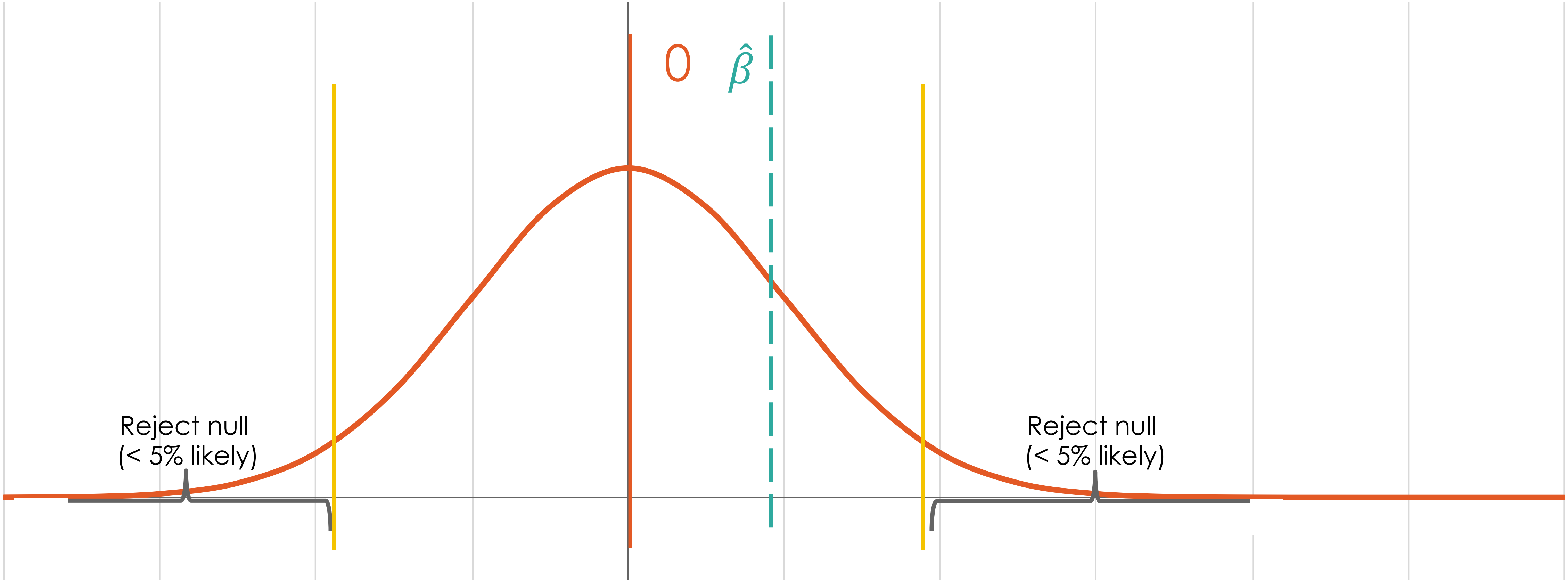
We say "$\hat{\beta}$ is statistically significantly different from zero at the 5% level" if an estimate this high (or low) has less than 5% probability under the null hypothesis.

Is $\hat{\beta}$ statistically significantly different from zero at the 5% level?



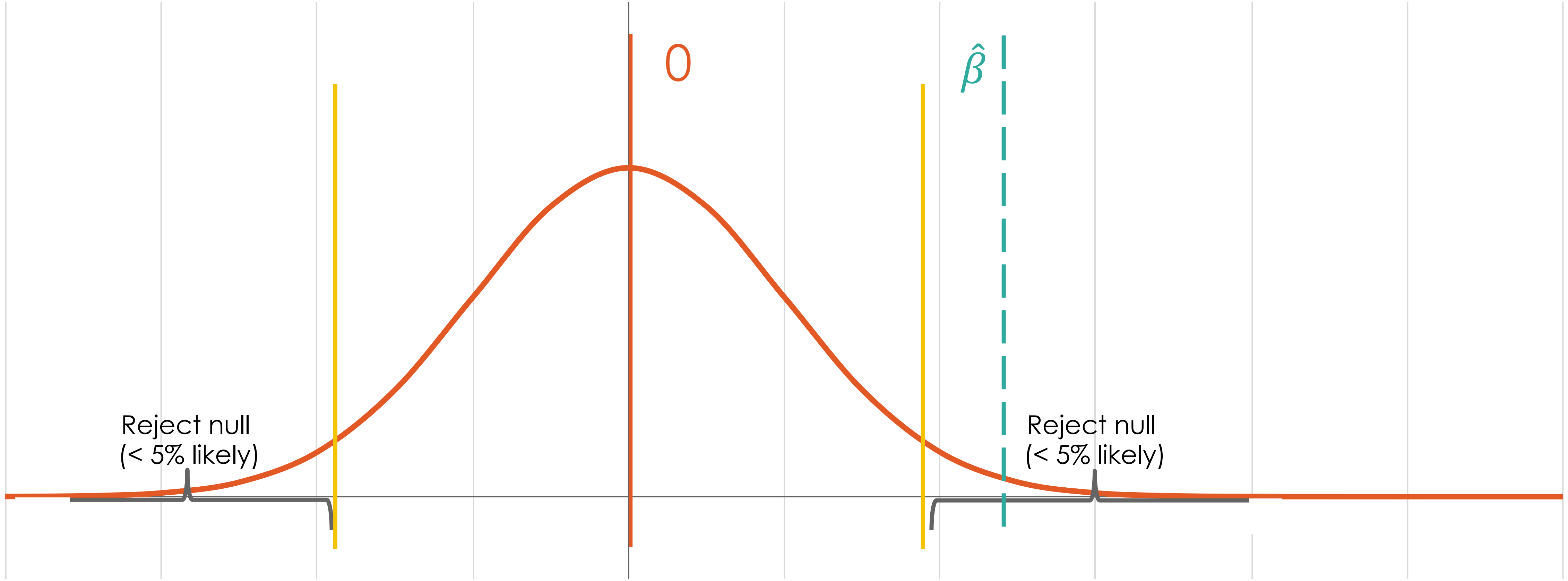
Only $\hat{\beta}$ that are large or small enough lead us to reject the null hypothesis of no treatment effect.

Is $\hat{\beta}$ statistically significantly different from zero at the 5% level?



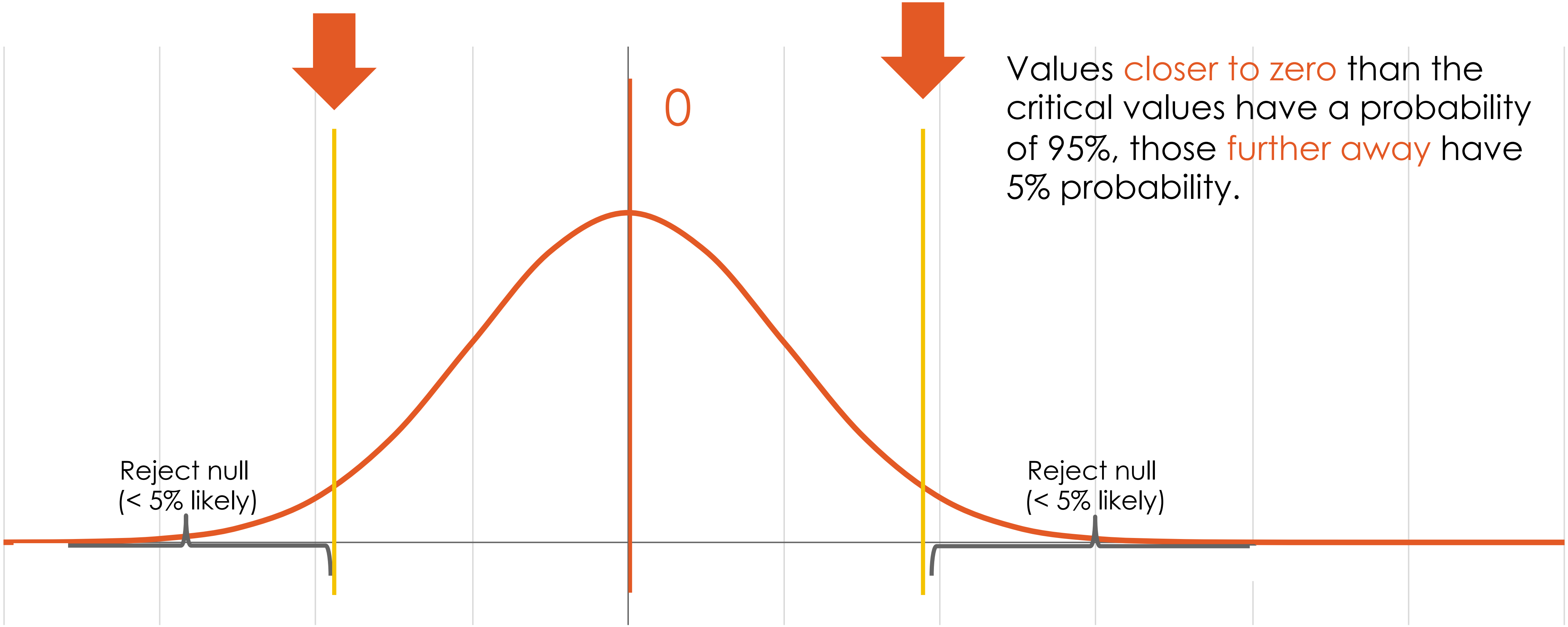
Only $\hat{\beta}$ that are large or small enough lead us to reject the null hypothesis of no treatment effect.

Is $\hat{\beta}$ statistically significantly different from zero at the 5% level?



Only $\hat{\beta}$ that are large or small enough lead us to reject the null hypothesis of no treatment effect.

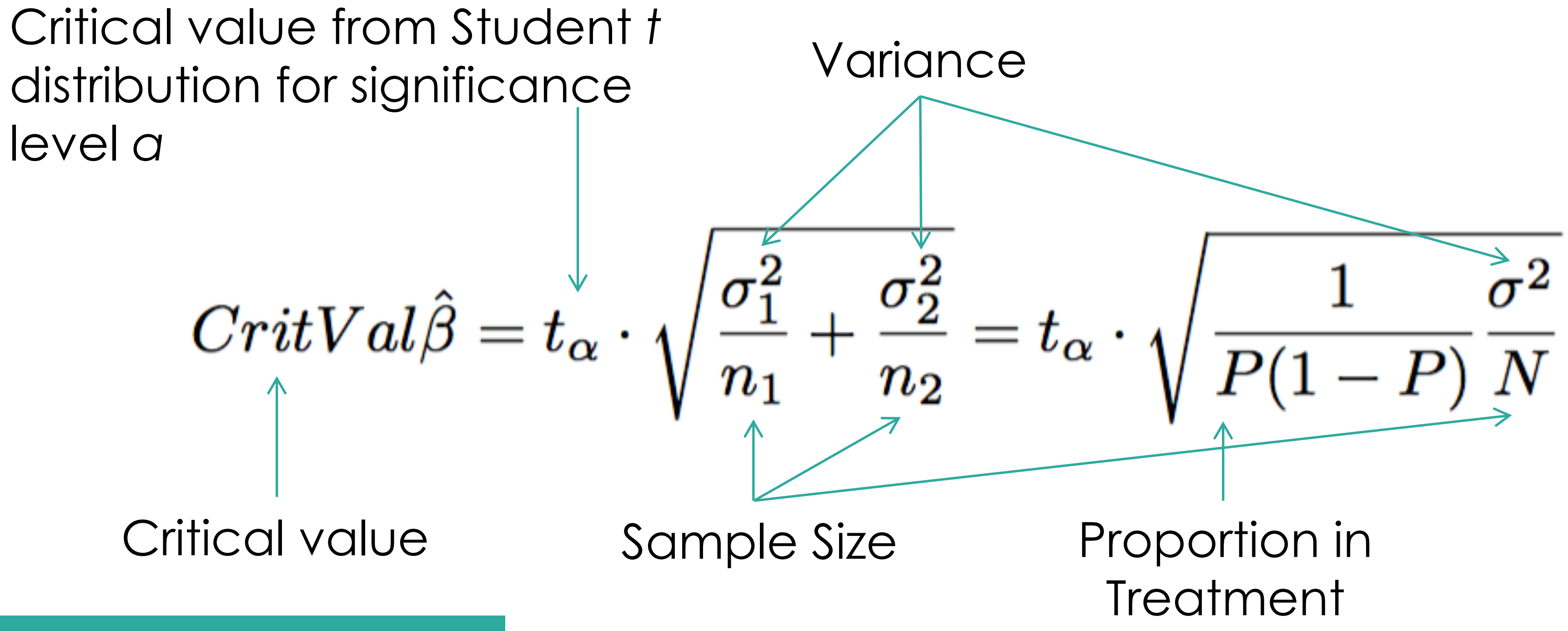
Critical Values for $\hat{\beta}$



Values closer to zero than the critical values have a probability of 95%, those further away have 5% probability.

The estimate $\hat{\beta}$ is statistically significant at the 5% level if it lies outside the critical values.

Critical Values for $\hat{\beta}$



Critical value from Student *t* distribution for significance level *α*

Variance

$$CritVal\hat{\beta} = t_{\alpha} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = t_{\alpha} \cdot \sqrt{\frac{1}{P(1-P)} \frac{\sigma^2}{N}}$$

Critical value

Sample Size

Proportion in Treatment

The critical value will be closer to zero with

- larger sample size *N*
- smaller outcome variance *σ*

Suppose the sample size is increased. For the same treatment estimate $\hat{\beta}$, is it more or less likely that the null is rejected?

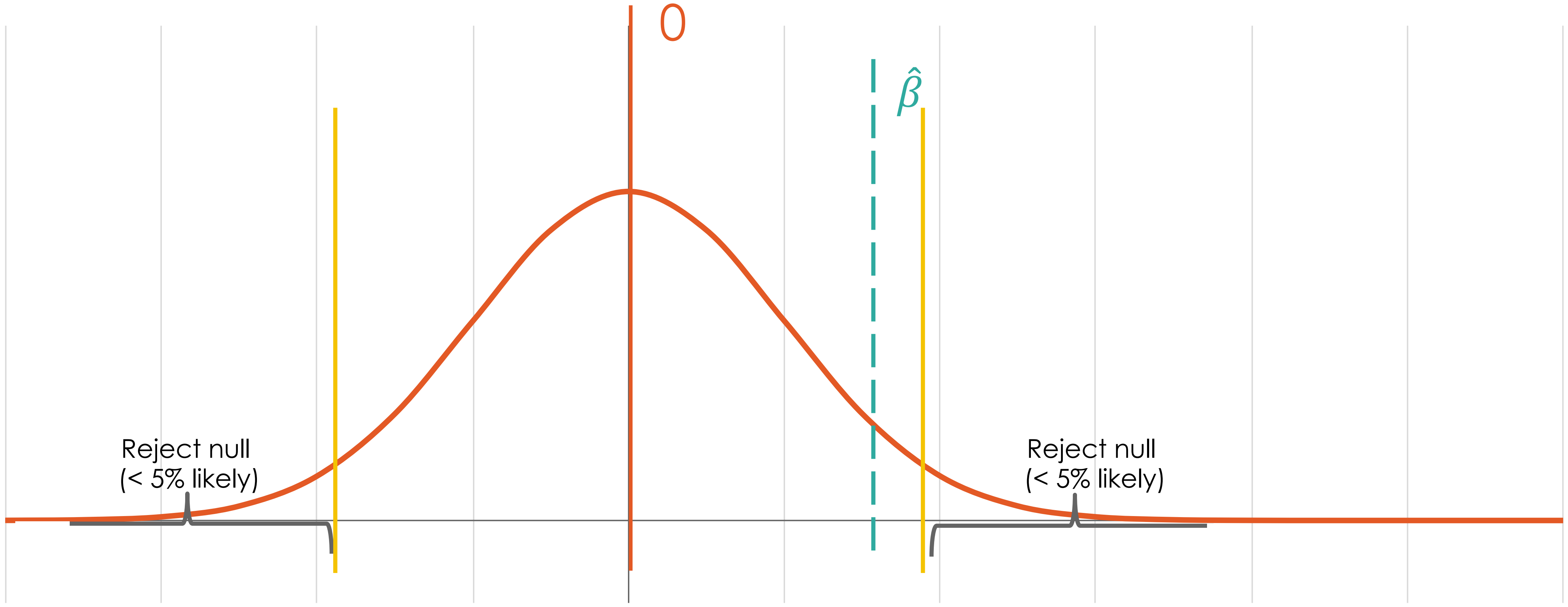
A. More likely

B. Less likely

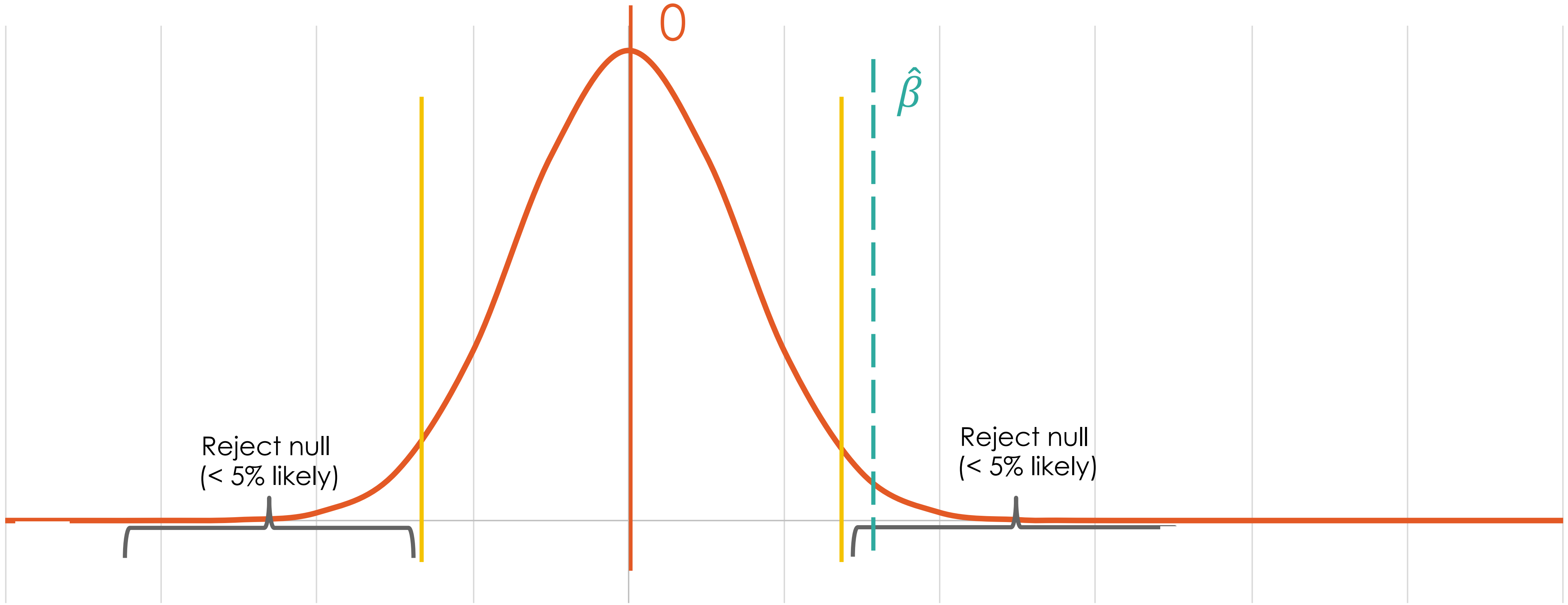
C. The probability is unaffected

D. Uncertain

An increase in the sample size



An increase in the sample size



Hypothesis Testing: Summary and Conclusions

- Larger treatment estimates $\hat{\beta}$ are less likely when the true treatment effect is zero. Randomization allows us to know *how* likely.

- If it is very unlikely (less than 5% probability) that the T-C difference is solely due to chance:
  - "We can reject the null hypothesis."
  - "The program has a statistically significant impact."

- This is a conservative approach.
- **Emphasis is on avoiding false positives (Type I error).**

Outline

I. True and Estimated Treatment Effect

II. Distribution of Treatment Effect Estimates

III. Hypothesis Testing

IV. Statistical Power and Sample Size

V. Power in Clustered Designs

VI. Calculating Power in Practice

Definition: Type II Error and Statistical Power

Type II error (false negative)

The probability of falsely concluding that there is no treatment effect.

The probability of not rejecting the null hypothesis (H_0), even though the null hypothesis is not true.

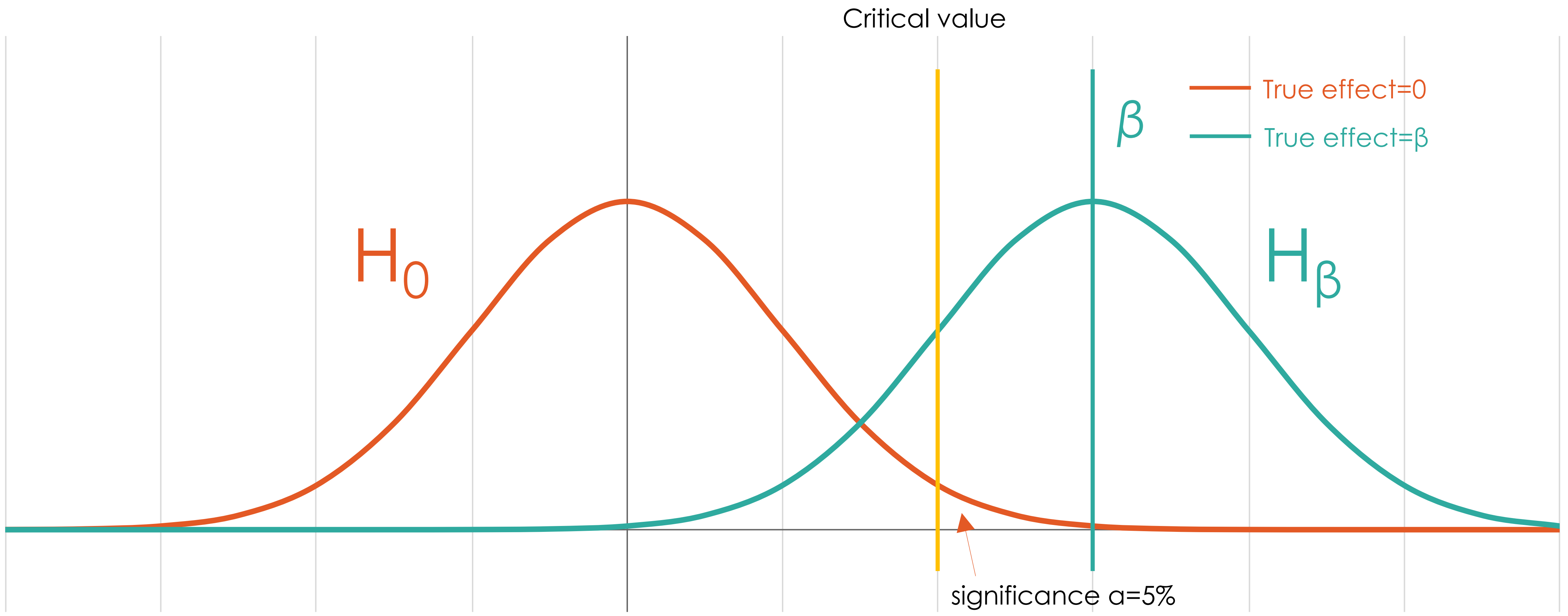
Statistical Power

The probability of a correct positive.

The probability of *avoiding* a Type II error.

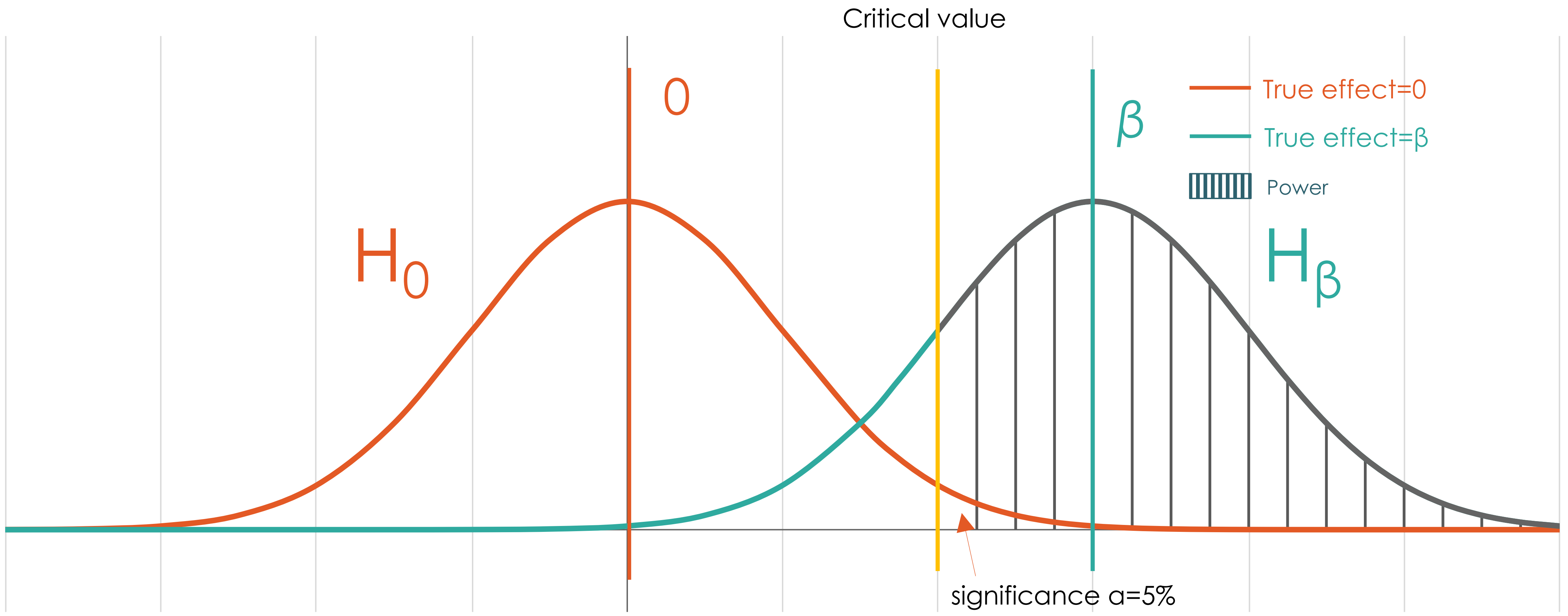
Typically, we aim for 80% power (some aim for 90%)

If the true effect was β…



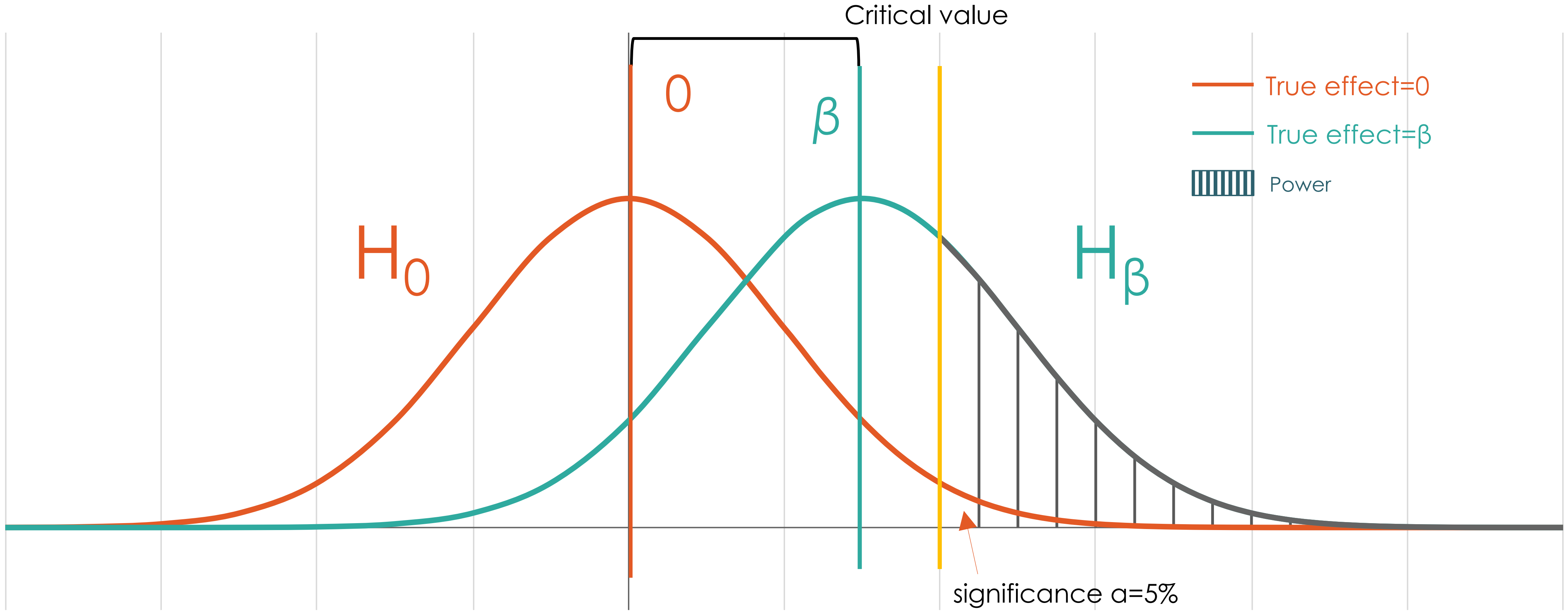
The distribution of treatment effect estimates is given by H_β.

… how often would we reject the null hypothesis?



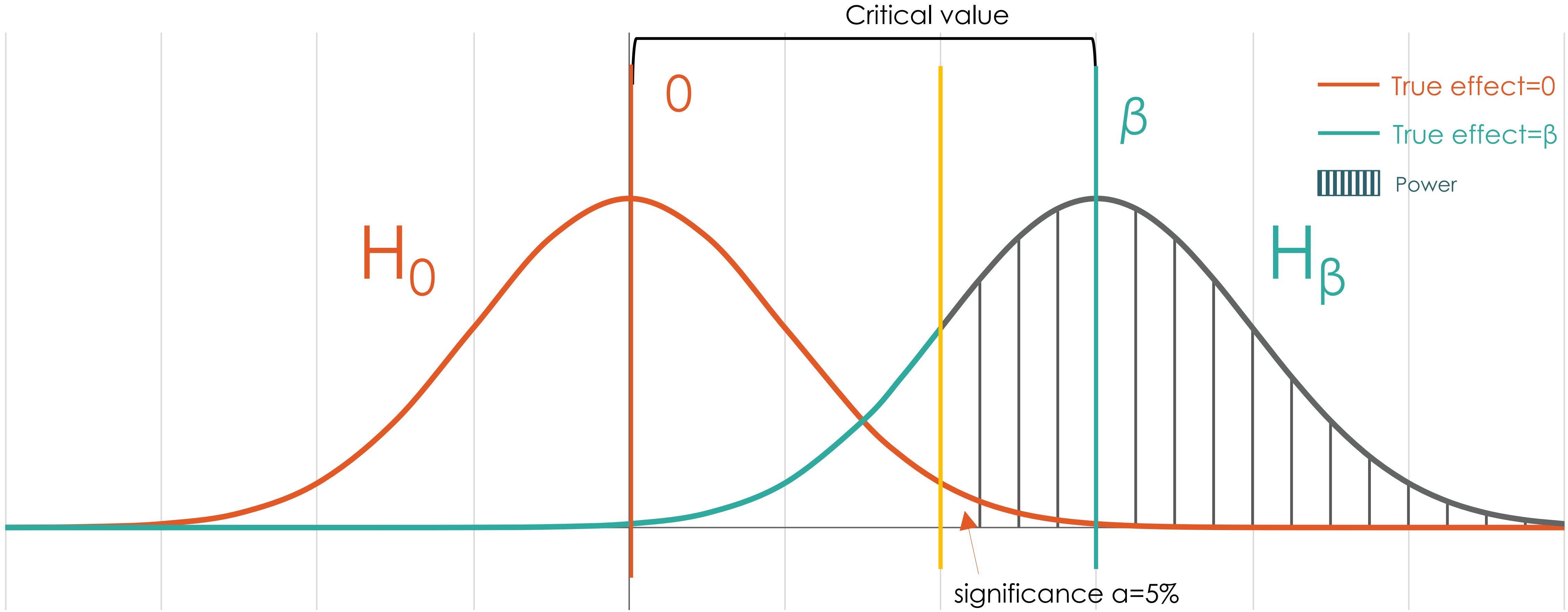
Shaded area shows power: probability that we obtain a treatment effect estimate which leads us to reject H_0 (when the true effect is β).

Power is lower with a smaller (true) treatment effect.



With a small β, the null hypothesis is rarely rejected.

Power is greater with a larger (true) treatment effect.



Greater effect size (relative to estimate dispersion) means more power

Power: Main Ingredients

1. Effect Size: a large effect is easier to distinguish from zero than a small effect.

Power: Main Ingredients

1. Effect Size: a large effect is easier to distinguish from zero than a small effect.

2. Variance: greater variability in the outcome variable makes it harder to distinguish an effect.

Power: Main Ingredients

1. Effect Size: a large effect is easier to distinguish from zero than a small effect.

2. Variance: greater variability in the outcome variable makes it harder to distinguish an effect.

3. **Sample Size**: a larger sample means that treatment and control are more representative of the overall population, making it easier to distinguish an effect.

Power: Main Ingredients

1. Effect Size: a large effect is easier to distinguish from zero than a small effect.

2. Variance: lower variability in the outcome variable makes it easier to distinguish an effect.

3. Sample Size: a larger sample means that treatment and control are more representative of the overall population, making it easier to distinguish an effect.

4. Sample split: an equal proportion in treatment and control makes it easiest to distinguish an effect.

Power: Main Ingredients

1. **Effect Size**: a large effect is easier to distinguish from zero than a small effect.

2. **Variance**: lower variability in the outcome variable makes it easier to distinguish an effect.

3. **Sample Size**: a larger sample means that treatment and control are more representative of the overall population, making it easier to distinguish an effect.

4. **Sample split**: an equal proportion in treatment and control makes it easiest to distinguish an effect.

Minimal Detectable Effect Size

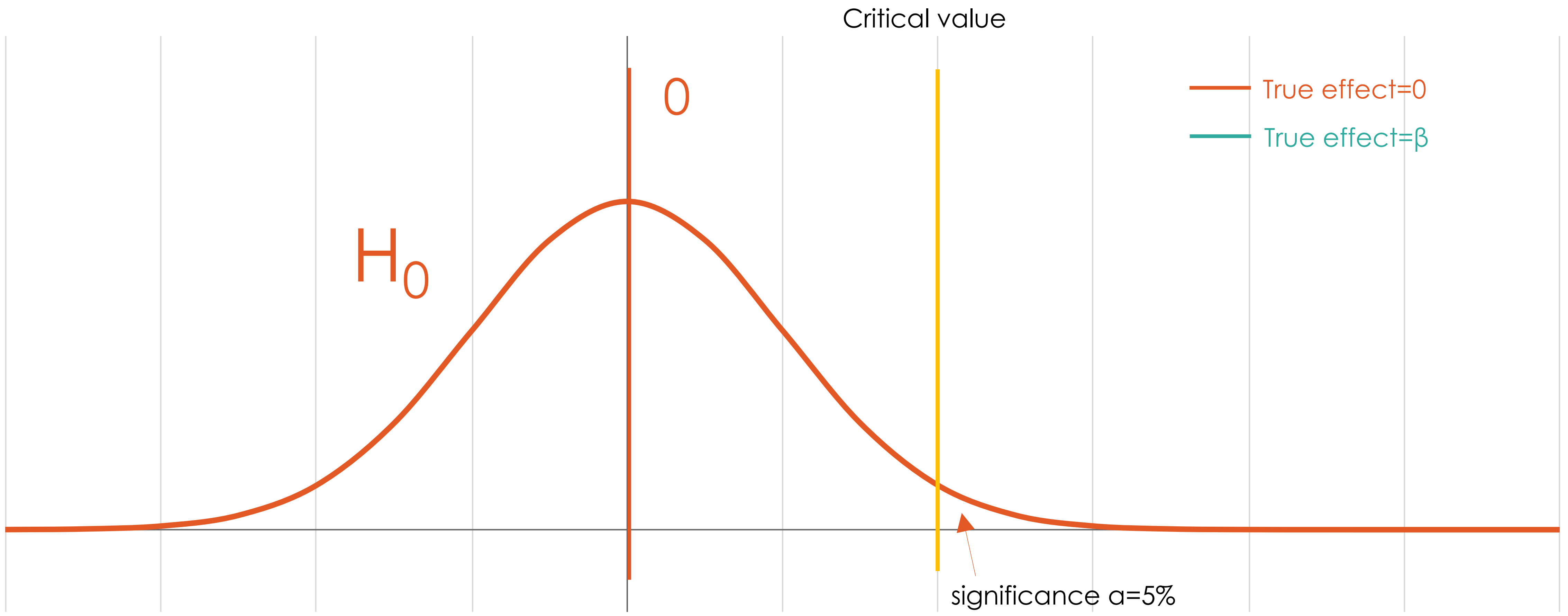
Minimal Detectable Effect (MDE)

The minimal effect size that can be detected with given statistical power (probability of correct positive, e.g. 80%), statistical significance (probability of a false positive, e.g. 5%) and sample size N.

- Ask: is it reasonable to expect effects as large or larger than the MDE?
- Would I like to be able to detect effects smaller than the MDE?

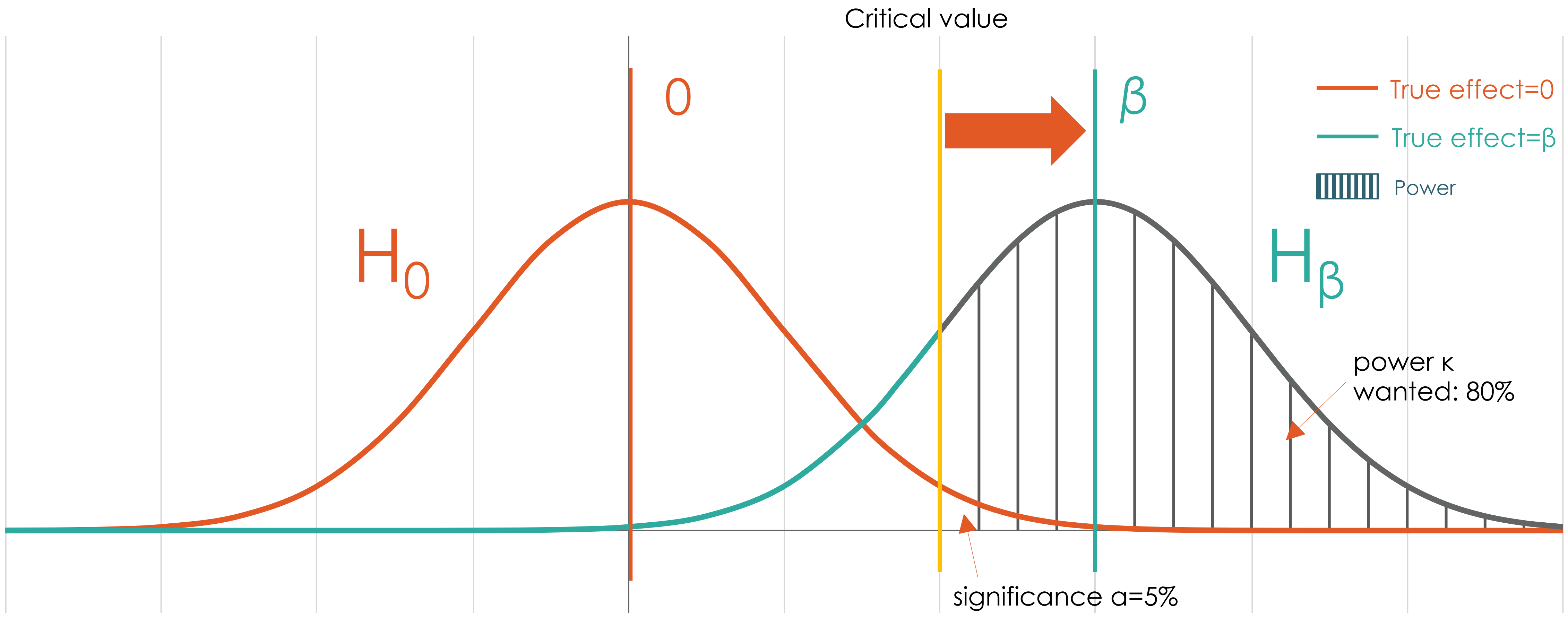
- Based on the MDE, can adjust the sample size to get to a realistic experimental design.

Minimal Detectable Effect Size (MDE)



How far out must β be so that we get κ = 80% power?

Minimal Detectable Effect Size (MDE)



How far out must β be so that we get κ = 80% power?

Minimal Detectable Effect Size (MDE)

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

Variance

$$MDE_\beta = (t_\kappa + t_\alpha) \cdot \sqrt{\frac{1}{P(1-P)} \cdot \frac{\sigma^2}{N}}$$

Minimal detectable effect

Proportion in Treatment

Sample Size

The MDE will be smaller with

- larger sample size *N*
- smaller outcome variance *σ*
- Even proportion in treatment (P = 0.5)

From a research design perspective, it can be justified to choose an uneven sample split if…

A. The treatment is likely to be highly effective

B. The number of people who can be treated is restricted

C. Additional data can be collected at low cost

D. B and C

E. Unsure

Outline

I. True and Estimated Treatment Effect

II. Distribution of Treatment Effect Estimates

III. Hypothesis Testing

IV. Statistical Power and Sample Size

V. Power in Clustered Designs

VI. Calculating Power in Practice

Definition: Intraclass Correlation

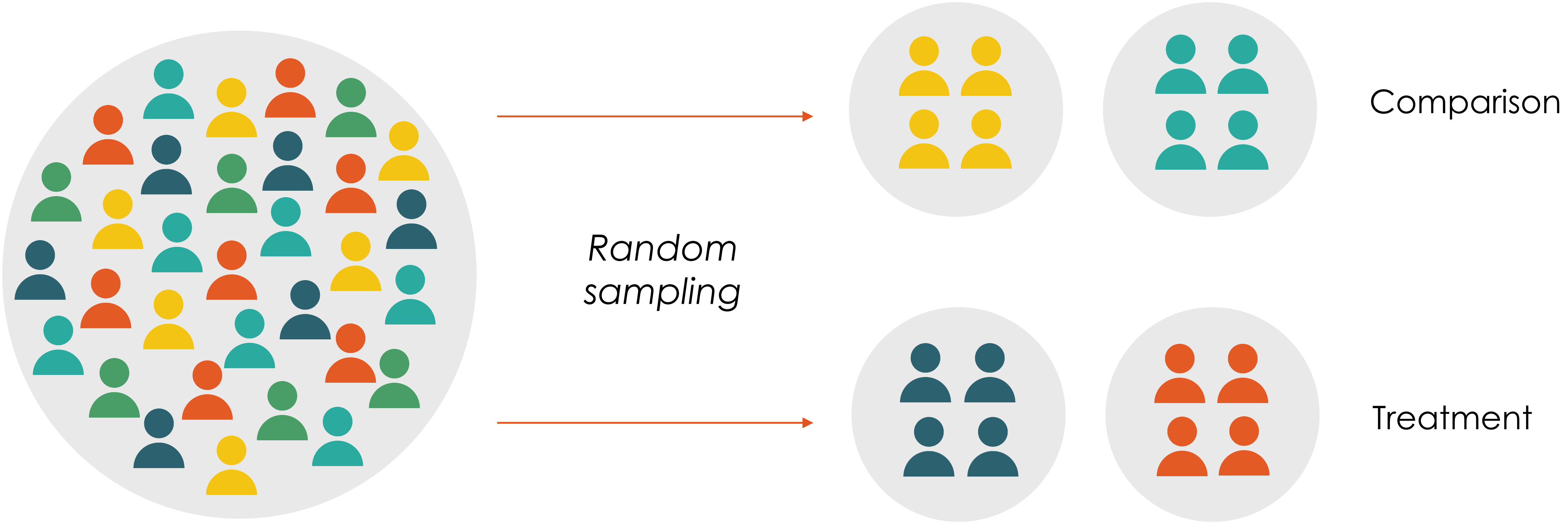
Intraclass correlation (ICC)

- The intraclass correlation describes how similar – how correlated – units within the same class or cluster are.
- The ICC is also the fraction of total variation accounted for by *between-class variation*.

- Total variance(σ^2) can be divided into within-class variance (σ_η^2), and between-class variance (σ_v^2).
- **High ICC (close to 1):** subjects *in the same cluster* are similar; different clusters tend to be very different from each other.
- **Low ICC (close to 0):** subjects in the same cluster are not particularly similar; different *clusters* tend to be similar to each other.

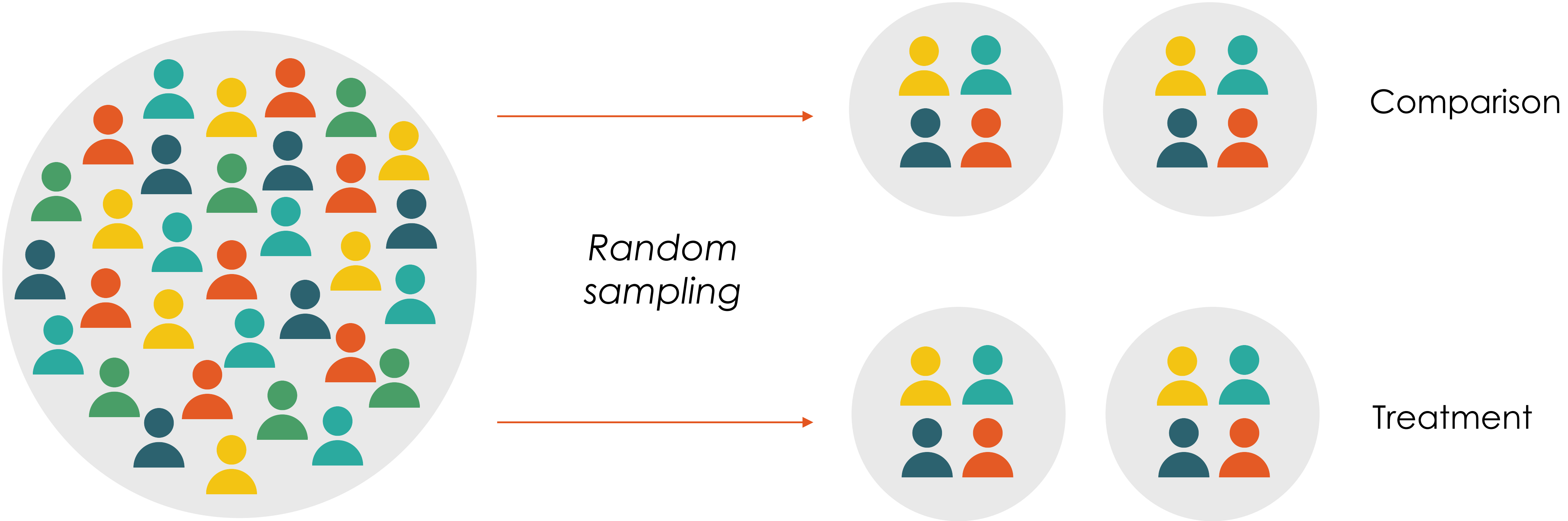
Intuition: how does the ICC affect power?

Samples with high intra-class correlation have similar individuals in each cluster, but different clusters are dissimilar.

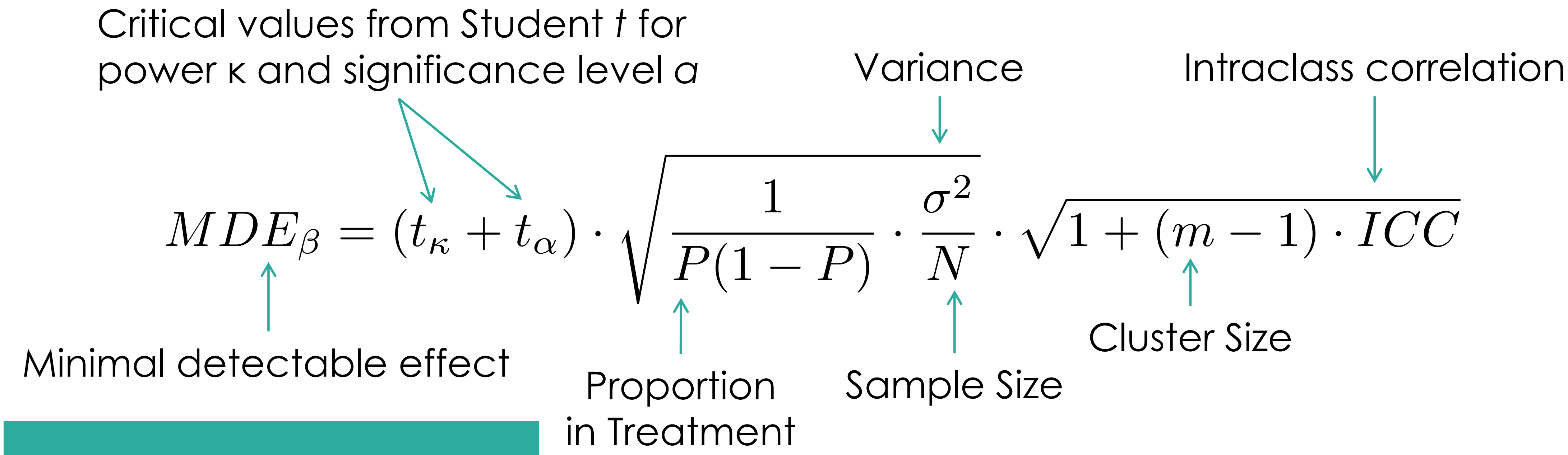


Intuition: how does the ICC affect power?

Samples with low intra-class correlation tend to have more similar clusters, and resemble the population as a whole more closely.



Minimal Detectable Effect Size (MDE)



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

Variance

Intraclass correlation

$$MDE_\beta = (t_\kappa + t_\alpha) \cdot \sqrt{\frac{1}{P(1-P)} \cdot \frac{\sigma^2}{N}} \cdot \sqrt{1 + (m-1) \cdot ICC}$$

Minimal detectable effect

Proportion in Treatment

Sample Size

Cluster Size

The MDE will be smaller with

- larger sample size *N*
- smaller outcome variance σ
- Even proportion in treatment (P = 0.5)
- lower ICC

Outline

I. True and Estimated Treatment Effect

II. Distribution of Treatment Effect Estimates

III. Hypothesis Testing

IV. Statistical Power and Sample Size

V. Power in Clustered Designs

VI. Calculating Power in Practice

Power calculations step by step

1. Set desired power (e.g. 80%) and significance level (e.g. 5%).
2. Decide allocation ratio (sample split), e.g. based on cost of data collection (control and treatment) and intervention (treatment only).
3. Ask: what treatment effect can be expected? What effect sizes would you like to be able to detect? → Use to set the MDE.
4. Estimate variance & ICC.
5. Back out the sample size and estimate the resulting study budget.

Estimating Variance and Intraclass Correlation

- MDE/ sample size for straightforward sampling designs can be calculated according to the formulas above.
  - *For more complex sampling designs or using additional data: use simulation.*

- Estimate within- and between-cluster variance of the outcome variable
  - *Check sensitivity of power calculations to different possible ICC values*

- **Where to find data on outcome variance?**
  - J-PAL and IPA DataVerses, World Bank Microdata Catalogue
  - National statistics
  - IPUMS or DHS data (large health and population household surveys)
  - Own data from program intervention

J-PAL
ABDUL LATIF JAMEEL POVERTY ACTION LAB

Thank you!

References, Reuse, and Citation

J-PAL, 2019