Development Learning Lab

Sampling and power

Viola Asri, Chr. Michelsen Institute

CMI CHR. MICHELSEN INSTITUTE UNIVERSITY OF BERGEN NHH SNF

Objective of this session

→ Understand the concept of sampling

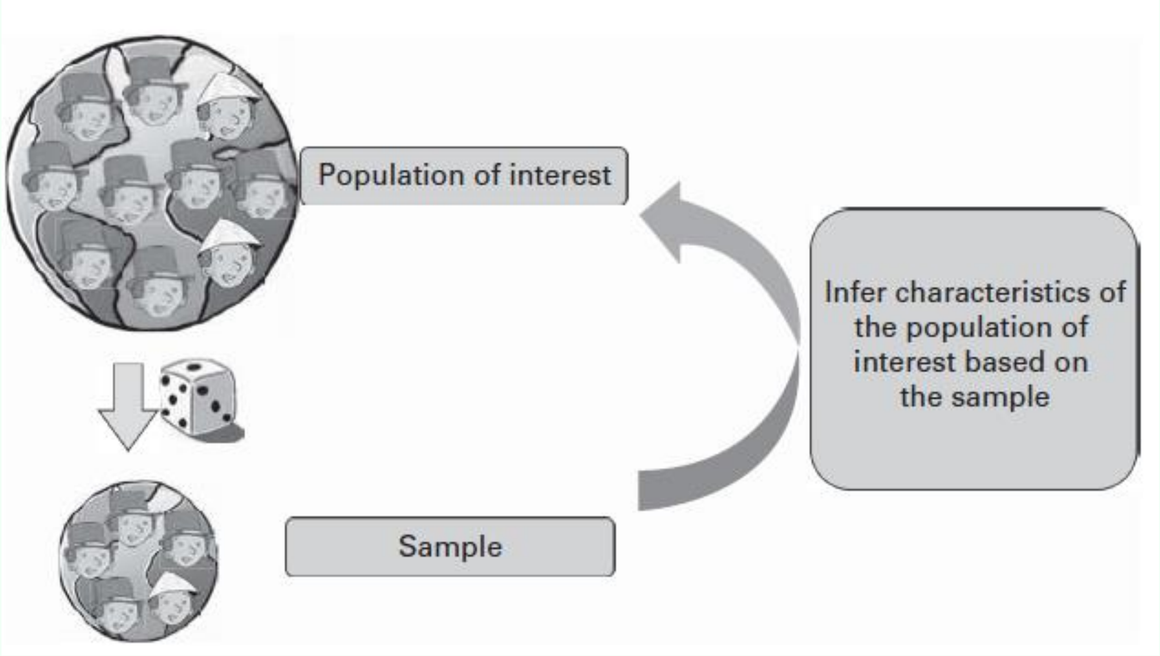
→ Intuitive familiarity with power calculations

How to do sampling?

Example: Measure impact of a nutrition program

→ What you would like to do but cannot? – Measure height for all children – „True effect“

→ What you need to do? – Sampling – Estimated effect



Source: Gertler et al. (2016), p. 262.

Step-by-step

1. Determine the population of interest
    a. Geographic area of the program
    b. Other characteristics: E. g. program eligibility
2. Identify a sampling frame – a comprehensive list of all units
3. Draw as many units from the sampling frame as required according to power calculations

Adapted from Gertler et al. (2016).

Sampling methods - probabilistic

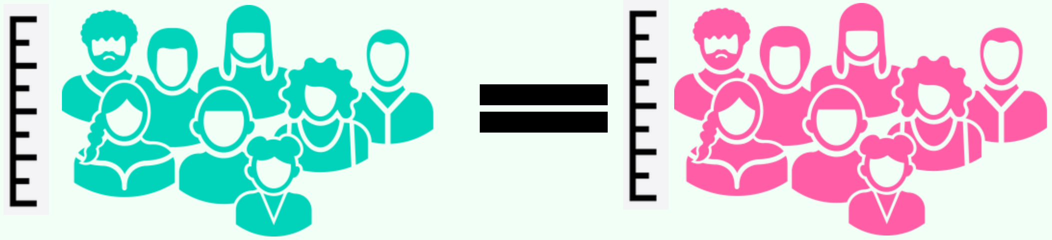
- Random sampling – every unit has the same probability
- Stratified random sampling – Population is divided into groups according to characteristics (e.g. male/female), random sampling within each group
- Clustered sampling – Units are grouped in clusters and a random sample of clusters is drawn
  - All units in a cluster constitute the sample or a random sample of the units in the cluster

Adapted from Gertler et al. (2016).

What sample size do we need?

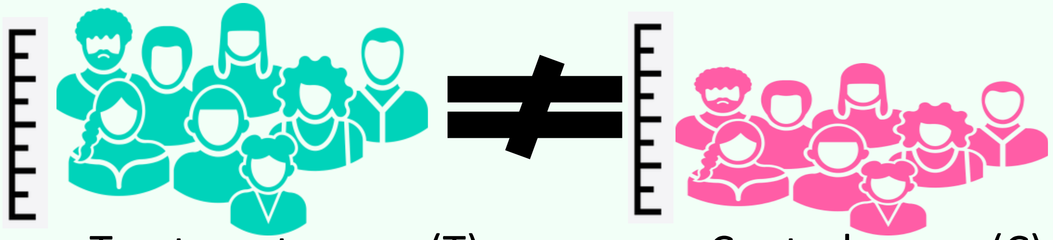
What do we test?

Treatment group (T) Control group (C)



Null hypothesis (H0) vs. Alternative hypothesis (H1)

„Presumption of zero“



Treatment group (T) Control group (C)

If the difference is unlikely (<5%) to be due to chance → Reject H0.

Adapted from Gertler et al. (2016).

Two potential errors

Type I

False positive: Detect an effect which in reality is not there. ➔ Wrongly reject H0.

Set significance level of 5% to limit the risk

Type II

False negative: Does not detect an effect but in reality it is there ➔ Wrongly keep H0.

Avoiding a false negative is POWER!

Adapted from Gertler et al. (2016).

Components of power calculations

Power calculations

Power = Probability that we can detect an effect if there is an effect.

→ Traditionally, we aim for 80% (or 90%)

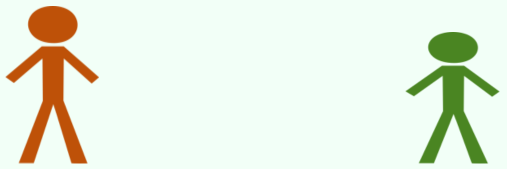
→ Main purpose: To determine the sample size that we need to detect an effect

→ Alternative: To determine the effect size we could detect given a sample size and other parameters

→ Why should we do our best to avoid low power?

Which effect size can we detect?

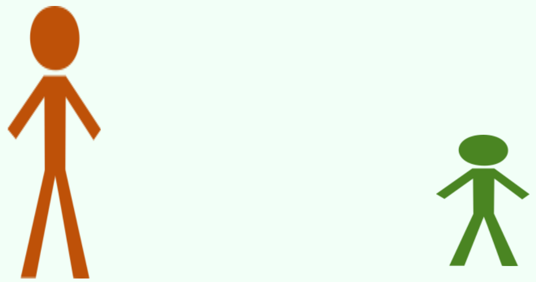
- Minimum detectable effect size: Smallest possible effect size that we will be able to detect.
- Typically set by the implementing organization or the researcher
  - Smallest effect for which it is worth it to continue or scale up the program ➔ Implementing organization‘s perspective
  - Smallest effect that you consider relevant to know ➔ Researcher‘s view



Treatment group average height

Control group average height

Small difference, harder to detect!

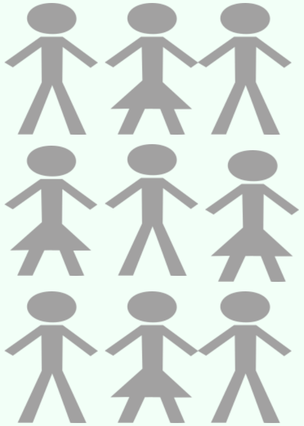


Treatment group average height

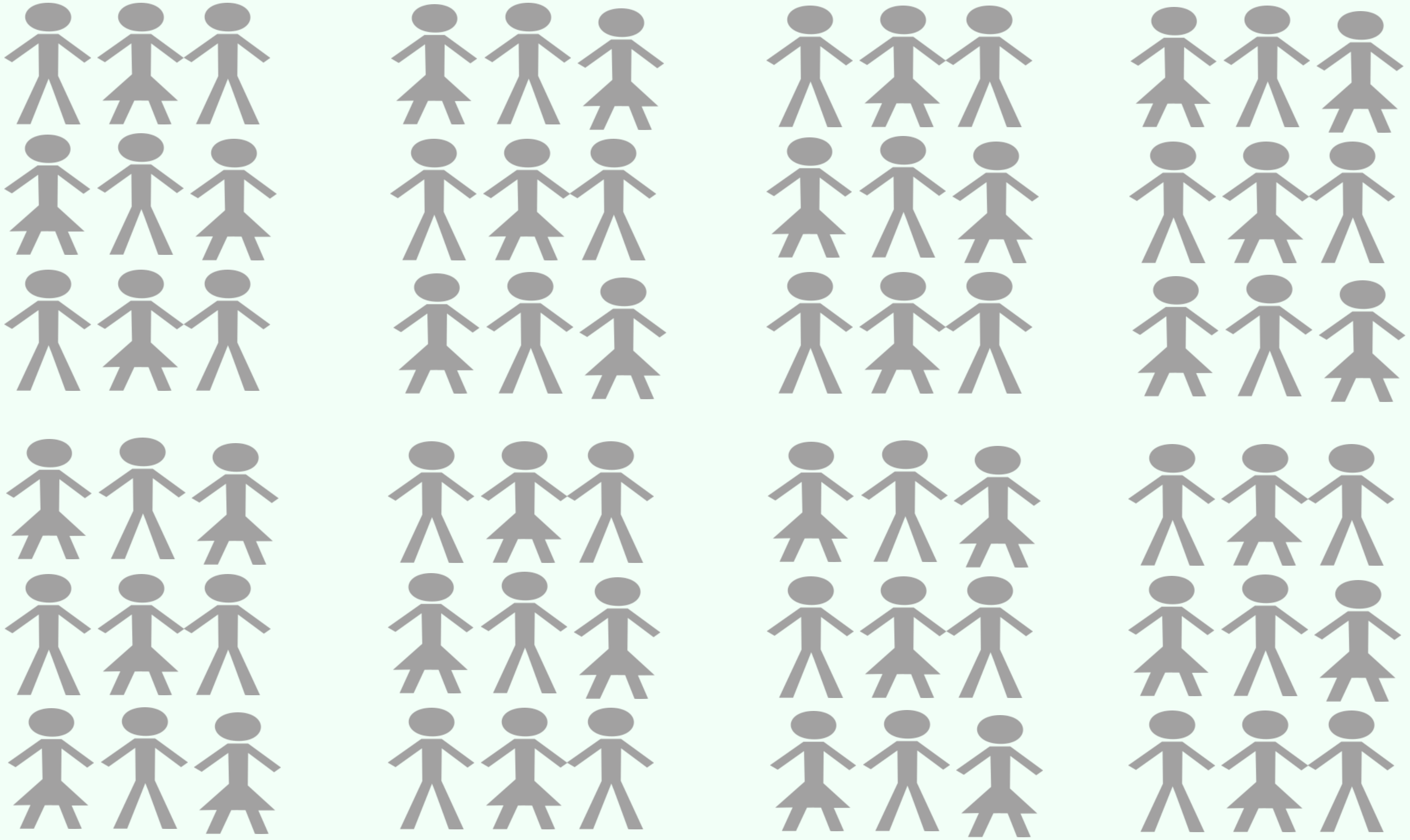
Control group average height

Large difference, easier to detect!

Higher sample, higher power?



vs.



Adapted from Martinez (2020)

Typically yes, but the level of treatment assignment matters.

Individual

Person, or household

Randomized controlled trial

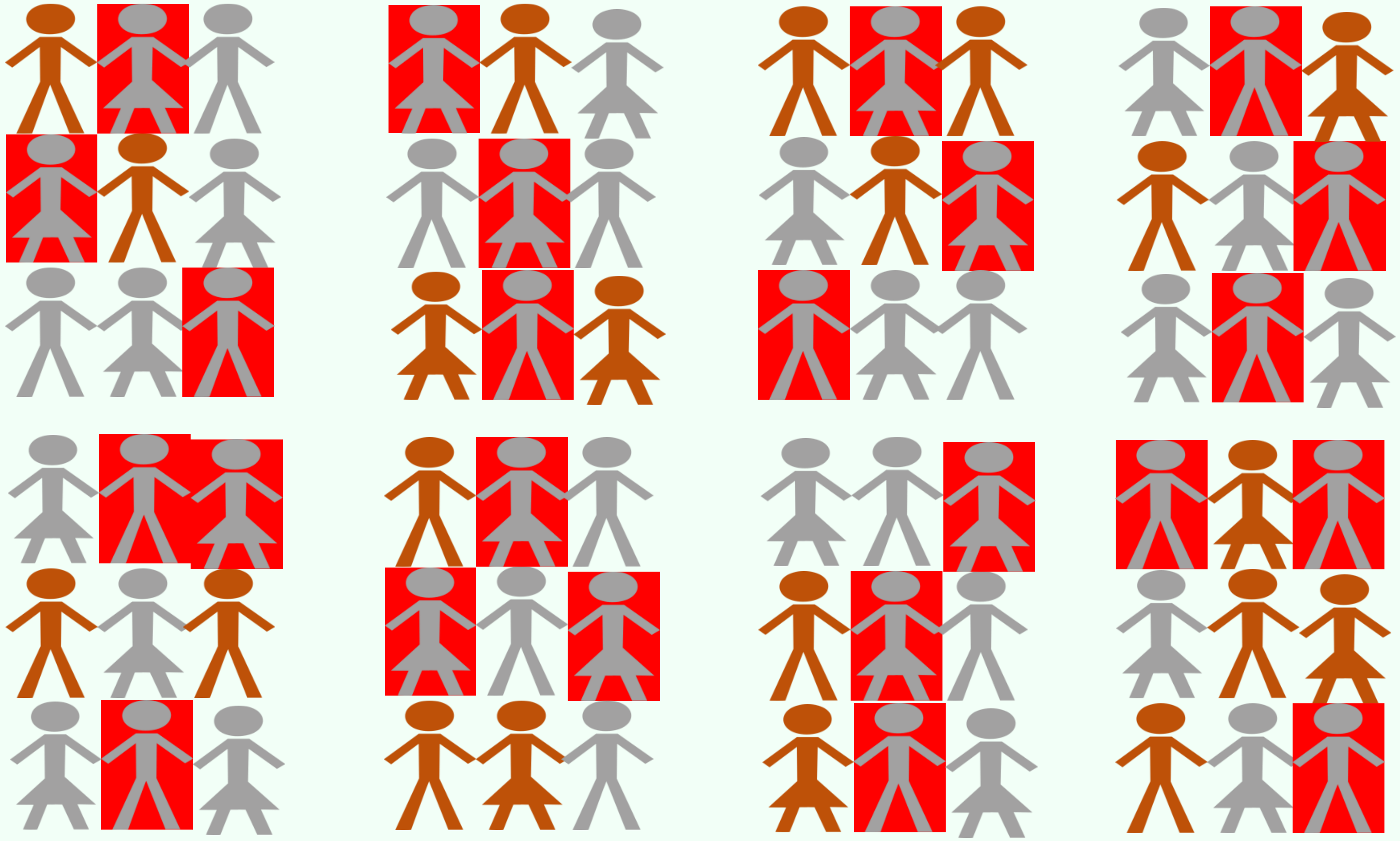
Cluster

Village, school, classroom, etc.

Clustered randomized controlled trial

Source: Martinez (2020)

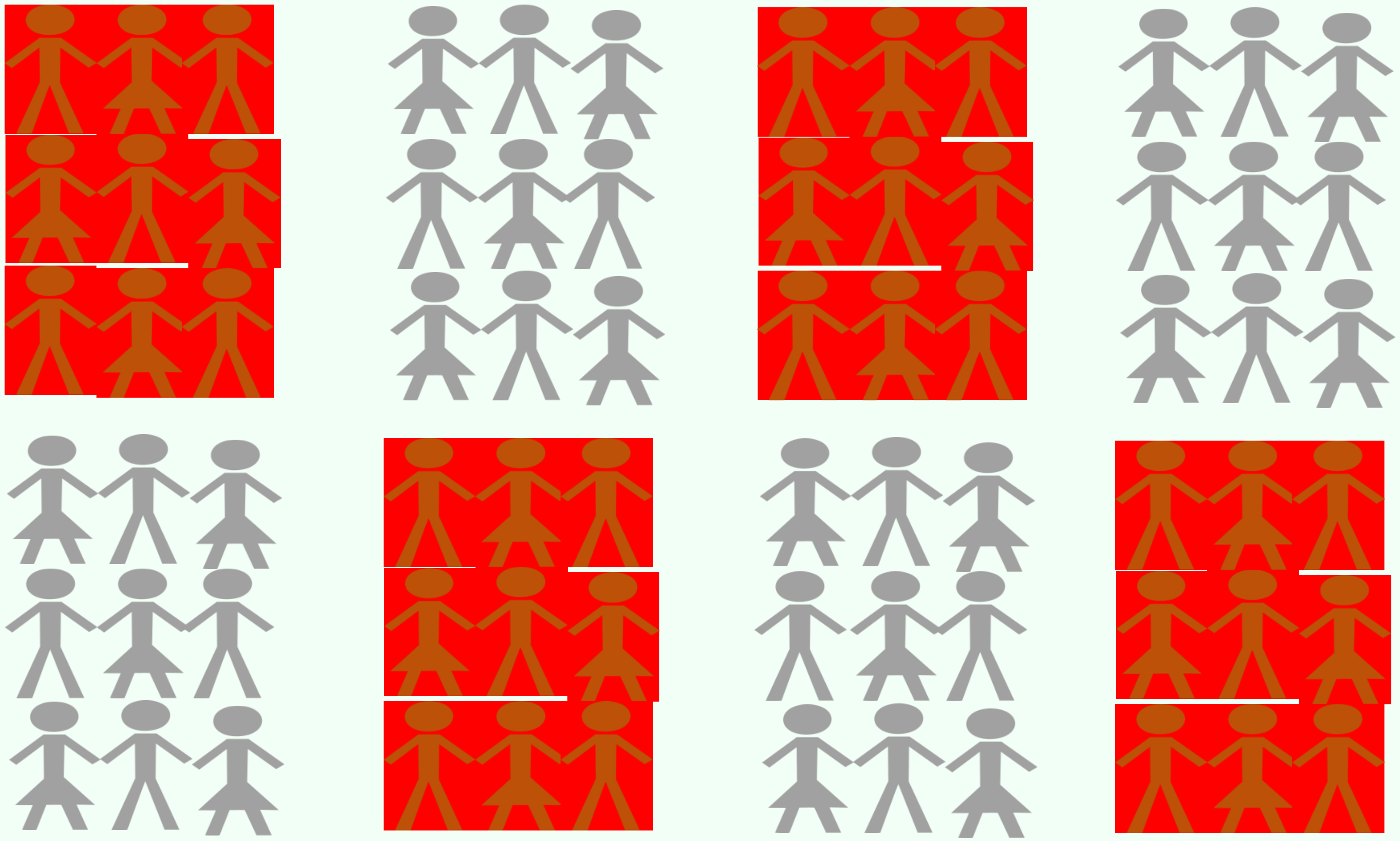
Individual assignment



Higher power but more potential for spillovers!

Source: Martinez (2020)

Cluster assignment



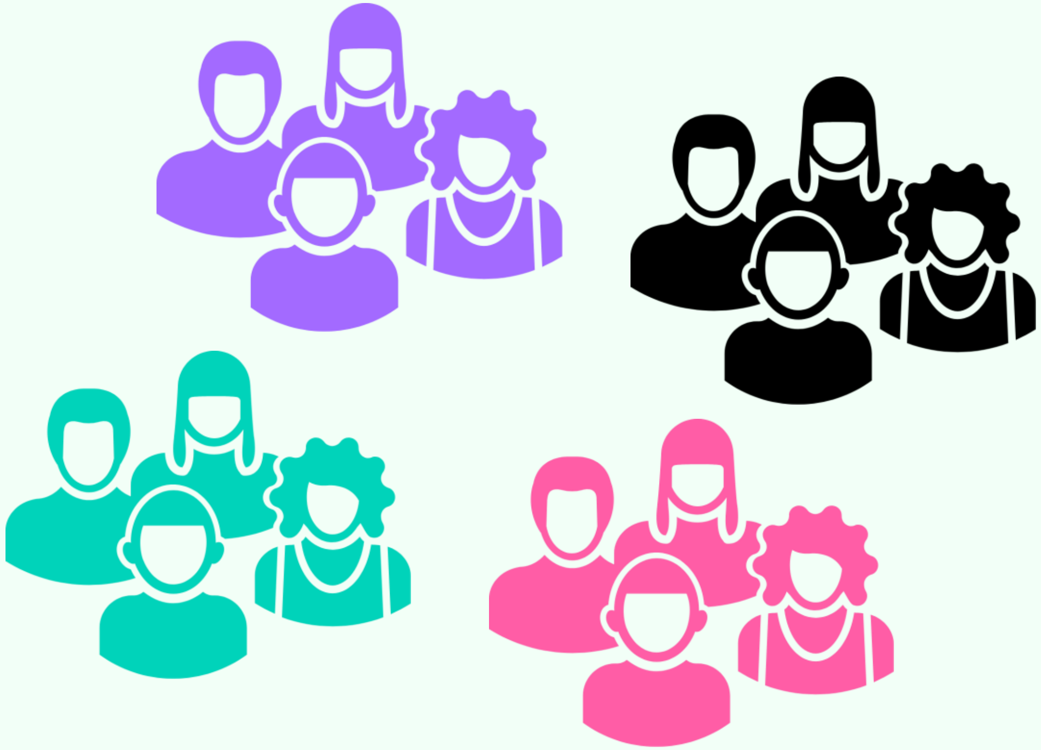
Less power (typically), less potential for spillovers/contamination

Source: Martinez (2020)

There are clusters … and clusters

Everyone is different in terms of the outcome of interest – low intra-cluster correlation!

Add more people from the clusters to increase power!



Everyone is the same in terms of our outcome of interest – high intra-cluster correlation!

Add more clusters to increase power!

Adapted from JPAL (2019)

Take-up rates / compliance

Only 50% of farmers participated in the training.

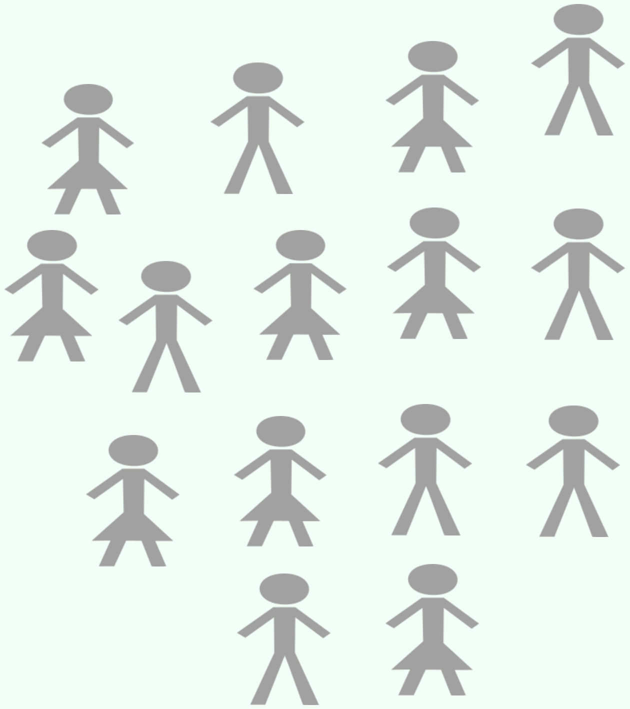
The effect is „diluted“ due to low take-up. ➔ Our power is reduced.

100% of the farmers participated in the training.

The effect is „complete“. ➔ No concern.

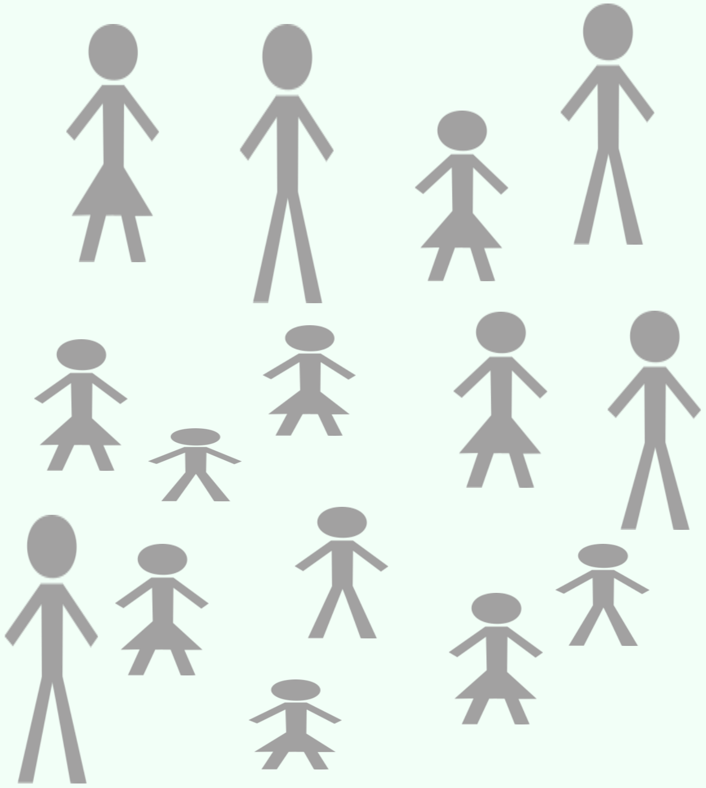
Adapted from JPAL (2019)

Variation in the outcome



All the kids are of similar height – all data points are closer together.

Higher power!



All kids are of very different height – all data points are very dispersed.

Lower power!

Adapted from JPAL (2019)

Summary

Factor	Relationship to power
Minimum detectable effect size ↓	↓
Sample size ↑	↑
Cluster vs. individual level	↓
Intra-cluster correlation ↑	↓
Take-up rates / compliance ↑	↑
Outcome variance ↑	↓

Adapted from Gupta and Kopper (2021)

Steps

- Decide about level of treatment
- Set desired power (80%, 90%) and significance (95%)
- Find data on your population and outcomes:
  - Calculate variance (and ICC if clustered)
- Set minimum detectable effect size
- Calculate sample size
- Estimate required budget and compare with implementation constraints!
- Adjust and repeat!

What can help to increase power?

- Collect baseline covariates – to reduce the outcome variance (e.g. baseline height of children)
- Improve take-up and compliance (compliance)
- Improve data quality – increases precision of your estimates

What do researchers actually do?

- → Use softwares that build on statistical formulas
- → Or simulate data collection process and power
- → Summarize relevant data for required parameters
- → Plug-in required parameters, using existing data such as
    - → Representative surveys
    - → Pilot data
    - → Administrative data
    - → Existing research (variance and intra-cluster correlation)

Group work

- → At what level is the program assigned? Individual or clustered?
- → What is your minimum detectable effect size?
- → What is the variance in your outcome of interest?
- → What is the intra-cluster correlation if clustered?
- → What sample size do you need? – Try out the tool (Excel template/EGAP) together with the expert!
- → Can you afford it?
- → ADJUST AND REPEAT!

Power calculation template (Excel)

For a prevalence

Language		1	(English)
Design parameters			
Power	$1 - \beta =$	80.0%	($t_{1-\beta}$ = 0.8416)
Confidence level	$1 - \alpha =$	95.0%	($t_{1-\alpha/2}$ = 1.9600)
Characteristics of the sample			
Cluster size	$m =$	1	For unclustered data, use m=1
Treatment clusters	$k_T =$	694	(n_T = 694)
Control clusters	$k_C =$	694	(n_C = 694)
Characteristics of the indicator			
Prevalence	$P =$	14.0%	(σ = 34.7%)
Minimum Detectable Effect	$\Delta =$	5.2%	

For a quantitative indicator (a continuous variable), click here

For an explanation of the formulas, click here

Power calculation template (Excel)

For a quantitative indicator (a continuous variable)

Language		1	(English)	
Design parameters				
Power	$1 - \beta =$	80.0%	$(t_{1-\beta} =$	0.8416)
Confidence level	$1 - \alpha =$	95.0%	$(t_{1-\alpha/2} =$	1.9600)
Characteristics of the sample				
Cluster size	$m =$	20	For unclustered data, use m=1	
Treatment clusters	$k_T =$	40	$(n_T =$	800)
Control clusters	$k_C =$	40	$(n_C =$	800)
Intra-Cluster Correlation	ICC =	0.1	(*Deff* =	2.9)
Characteristics of the indicator				
Standard deviation	$\sigma =$	1.00		
Minimum Detectable Effect	$\Delta =$	0.24		

For a prevalence, click here

For an explanation of the formulas, click here

Other tools

- For individual-level design: Interactive visualization https://rpsychologist.com/d3/nhst/
- EGAP Power Calculator: https://egap.shinyapps.io/power-app/

References and further reading

Gertler, P. J., Martinez, S., Premand, P., Rawlings, L. B., & Vermeersch, C. M. (2016). *Impact evaluation in practice*. World Bank Publications, available at https://openknowledge.worldbank.org/entities/publication/ebbe3565-69ff-5fe2-b65d-11329cf45293.

Gupta, S. & Kopper, S. (2021). Power calculations available at https://www.povertyactionlab.org/resource/power-calculations.

JPAL (2019). Statistical Power and Choosing the Right Sample Size available at https://www.povertyactionlab.org/sites/default/files/research-resources/L5ChoosingTheRightSampleSize.pdf.

Martinez, S. (2020). Statistical Power. Presentation.

Naimpally, R. & Wack, B. (2018). Six rules of thumb for understanding statistical power available at https://www.povertyactionlab.org/blog/5-21-18/six-rules-thumb-understanding-statistical-power.

Development
Learning
Lab

CMI CHR. MICHELSEN INSTITUTE

UNIVERSITY OF BERGEN

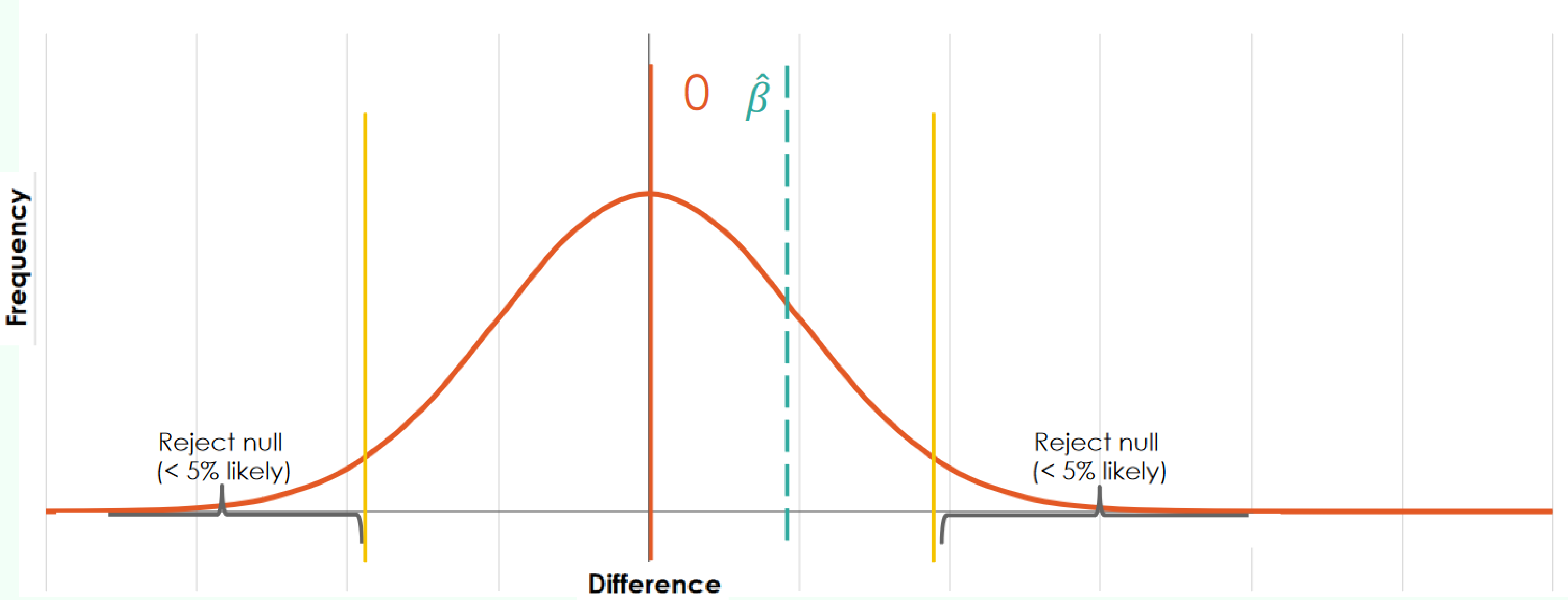
NHH

SNF

Statistical illustrations

Is the estimated effect ($\hat{\beta}$) significantly different from zero at the 5% level?

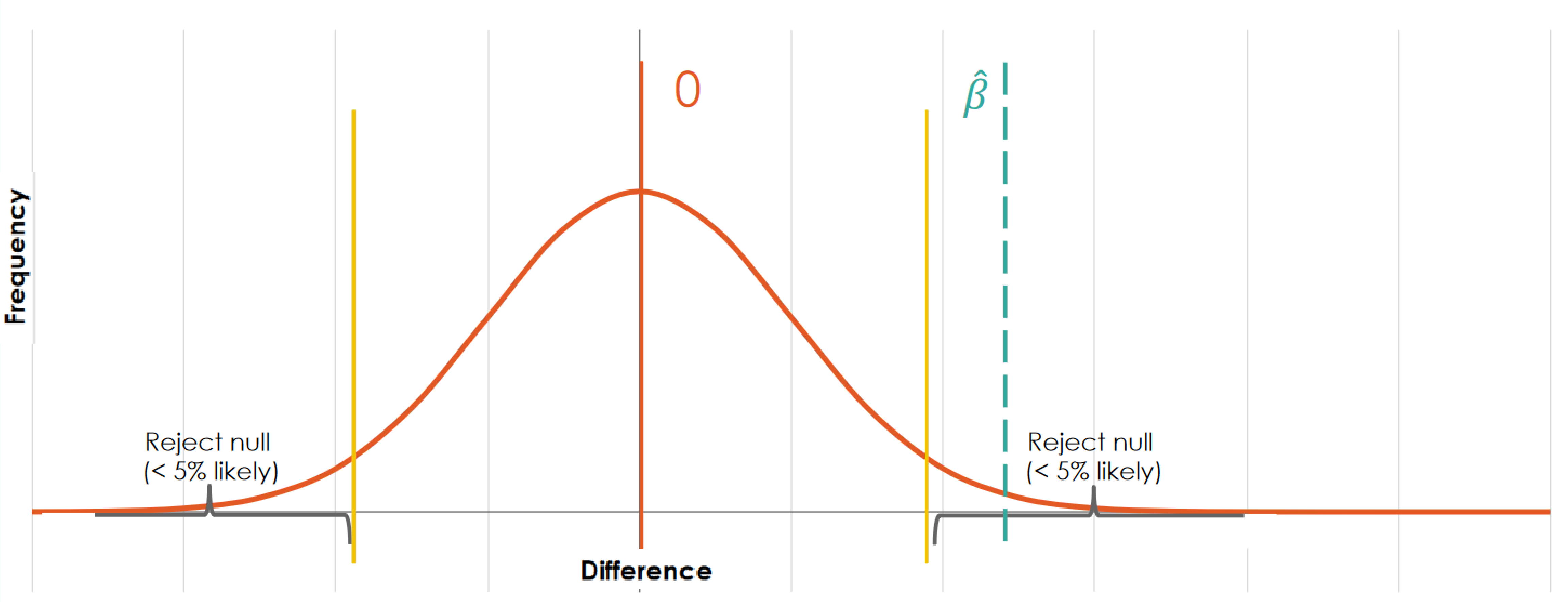
Ask: how likely is it that we would see an estimate as large as in an experiment, if the true effect ß was actually zero?



We cannot reject the null-hypothesis as the estimated effect is too close to 0 – it might have been due to chance.

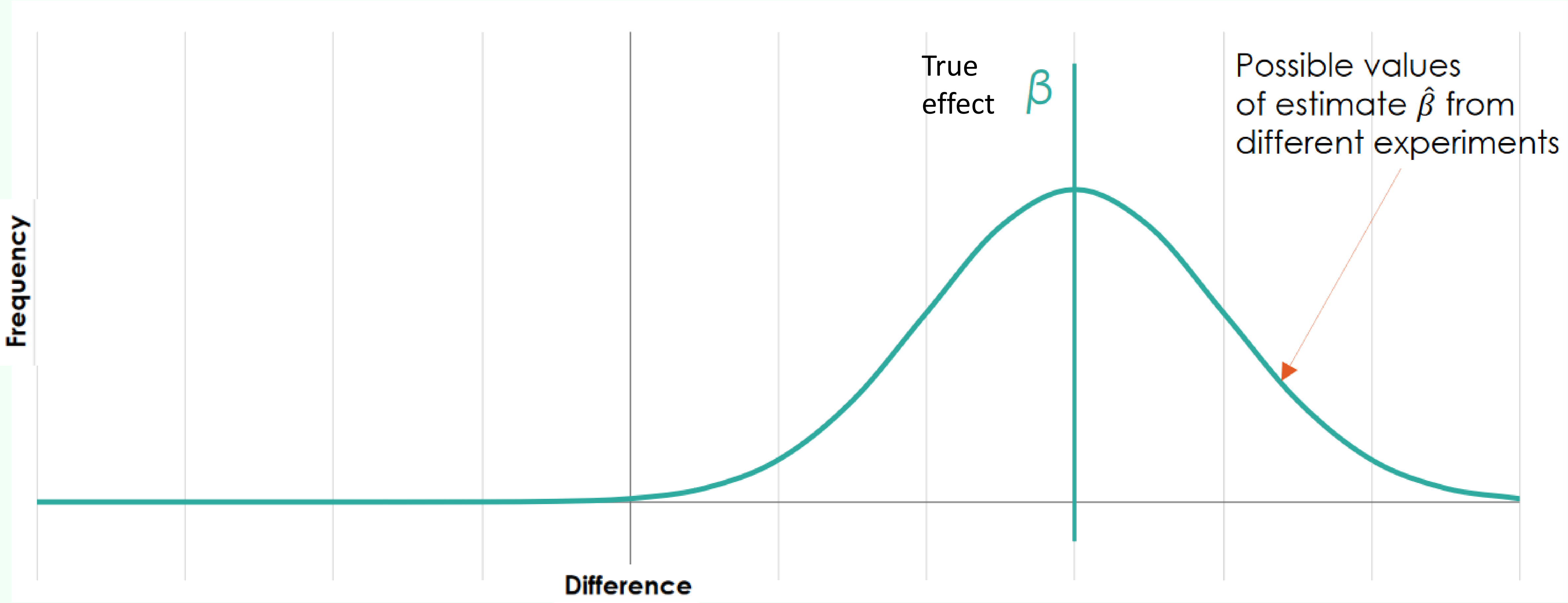
Is the estimated effect ($\hat{\beta}$) significantly different from zero at the 5% level?

Ask: how likely is it that we would see an estimate as large as in an experiment, if the true effect was actually zero?

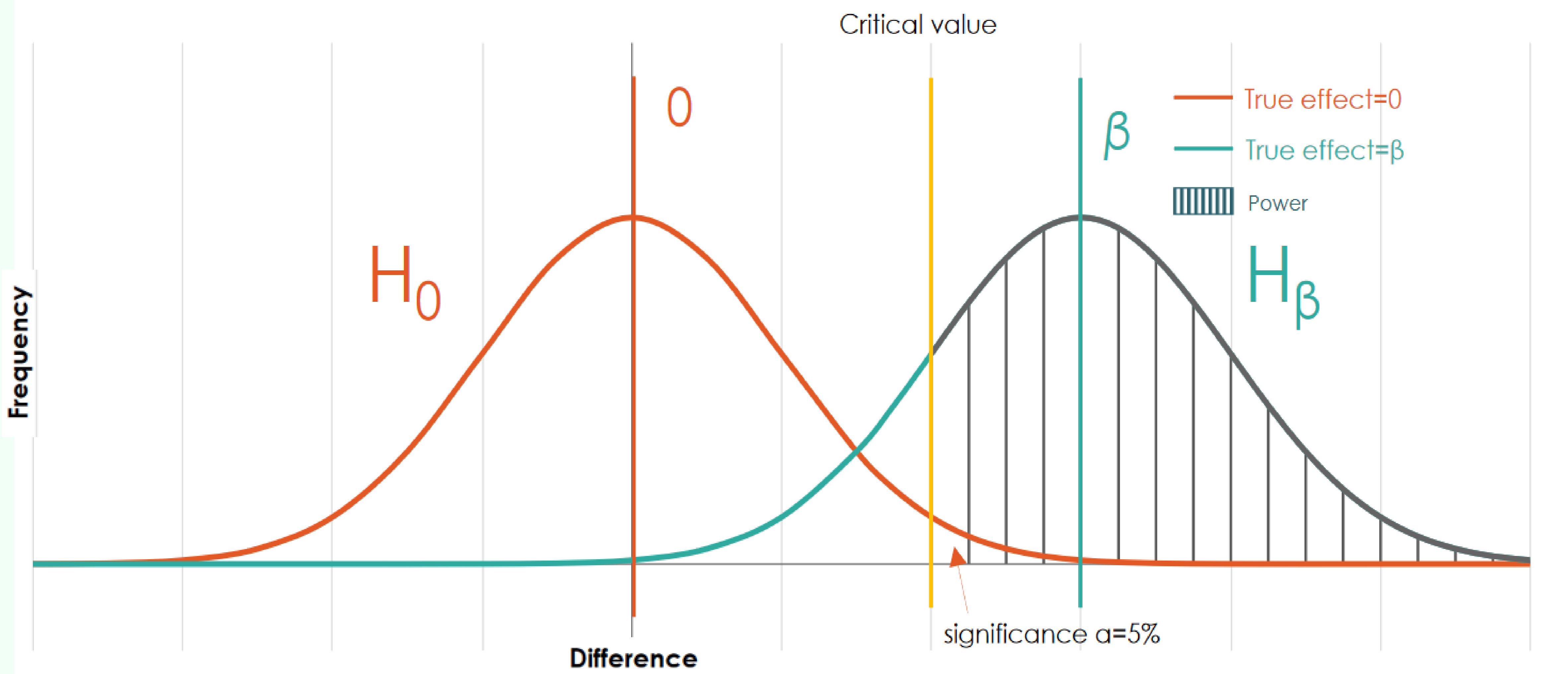


We reject the null-hypothesis as the estimated effect is large and unlikely due to chance. Less than 5%.

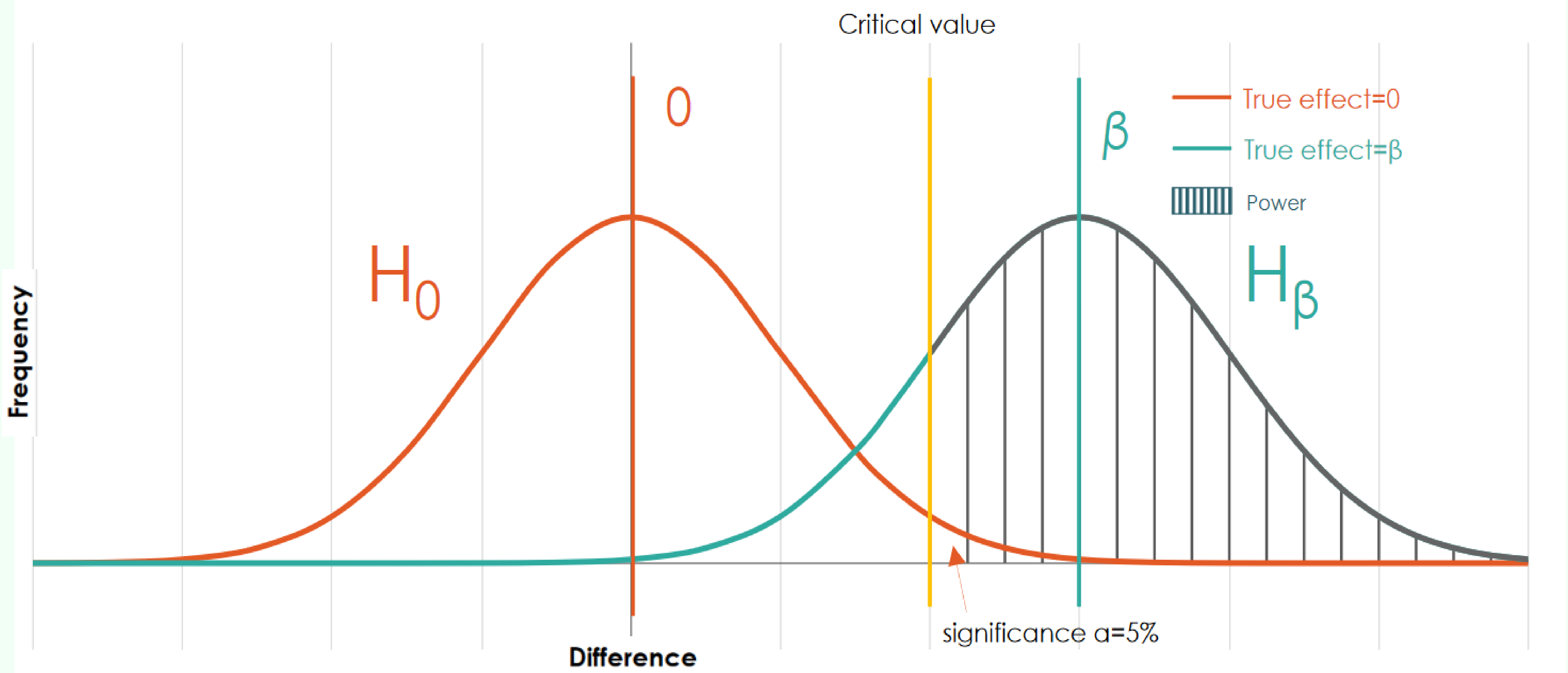
Imagine you could run an RCT an indefinite number of times



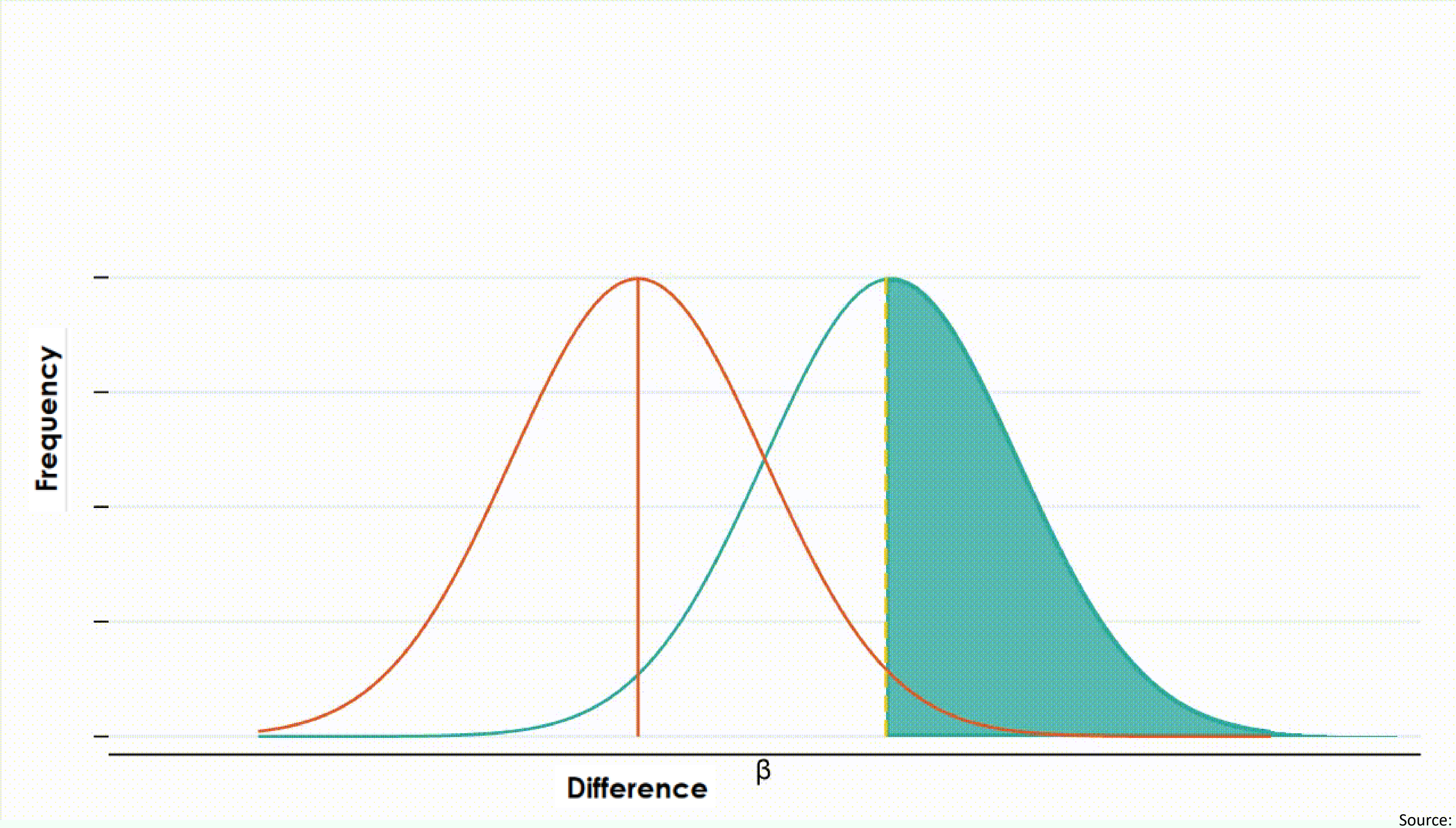
A statistical visualization of power



A statistical visualization of power

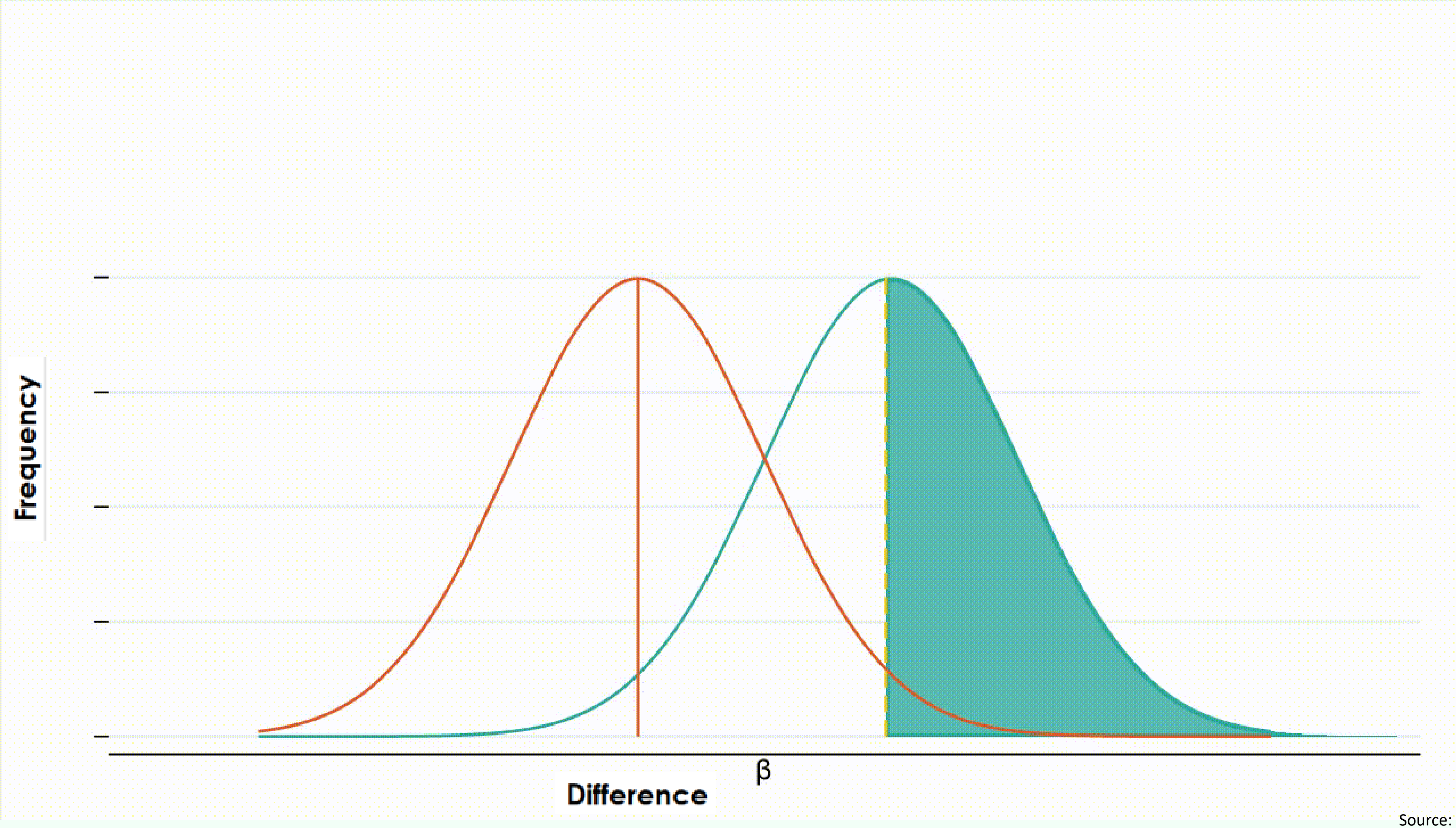


Sample size ↑ => Power ↑



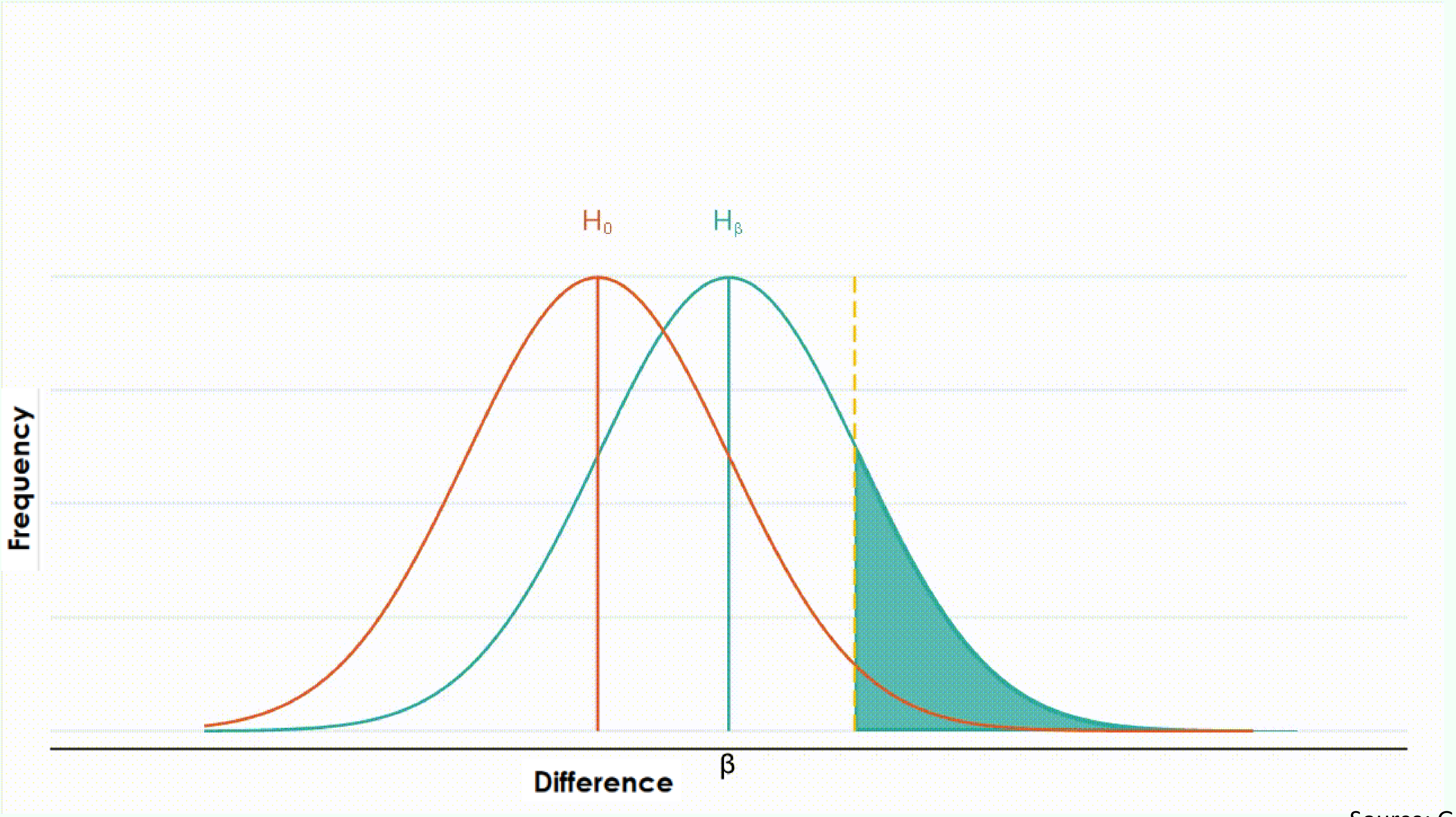
Source: Gupta and Kopper (2021)

Outcome variance ↓ => Power ↑



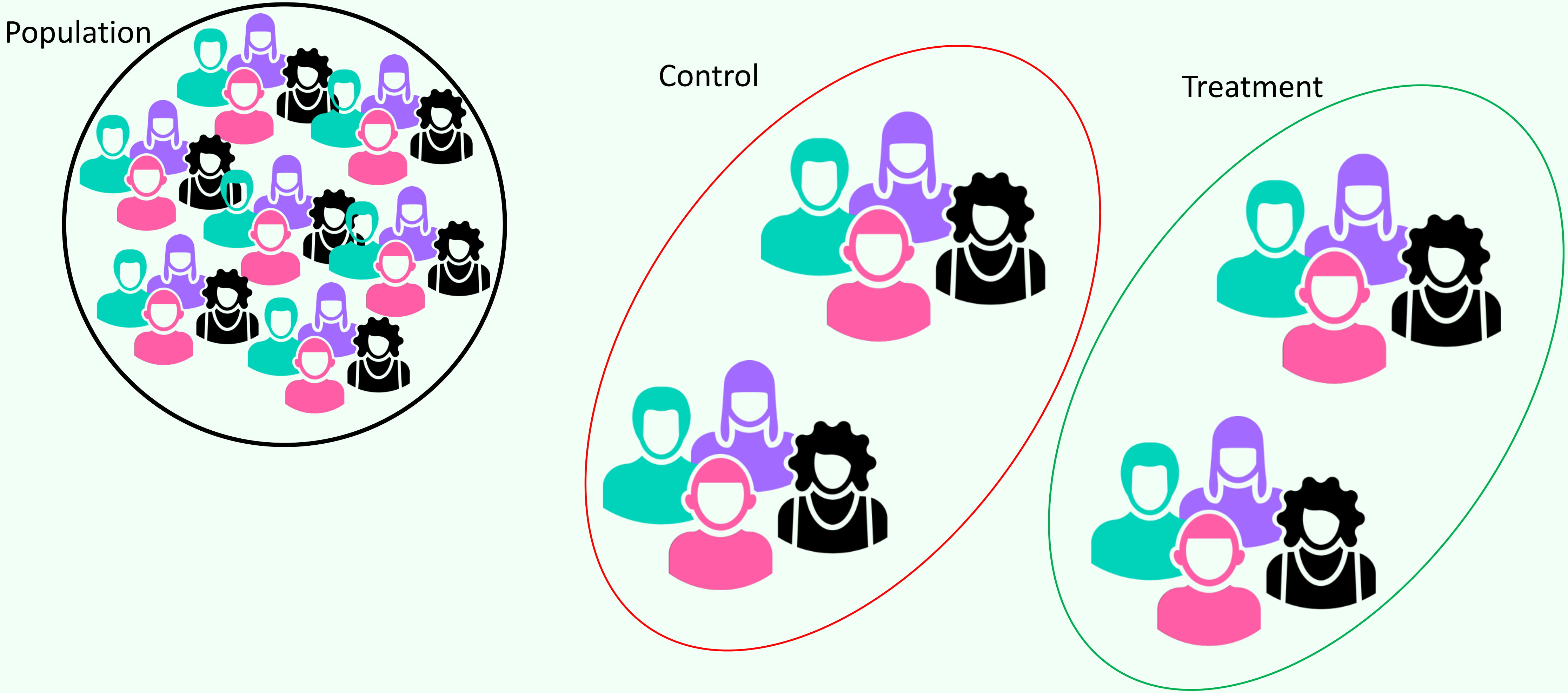
Source: Gupta and Kopper (2021)

Take-up rate / compliance ↑ => Power ↑

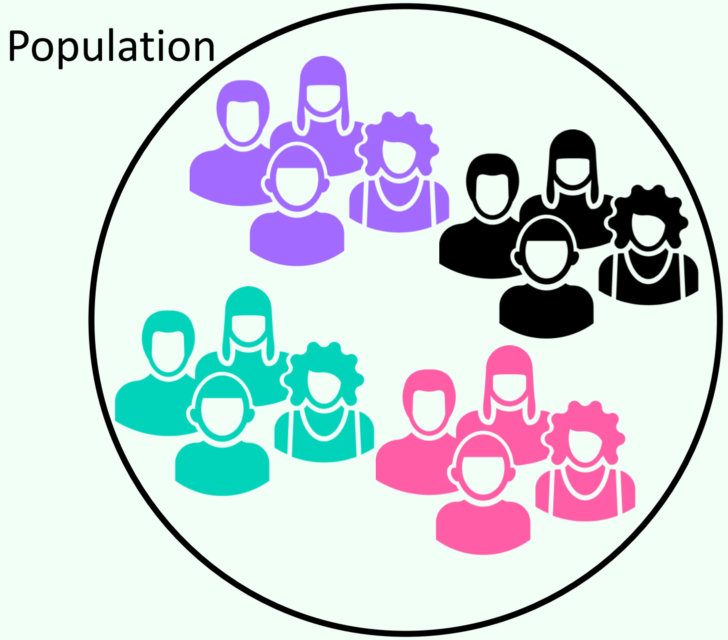


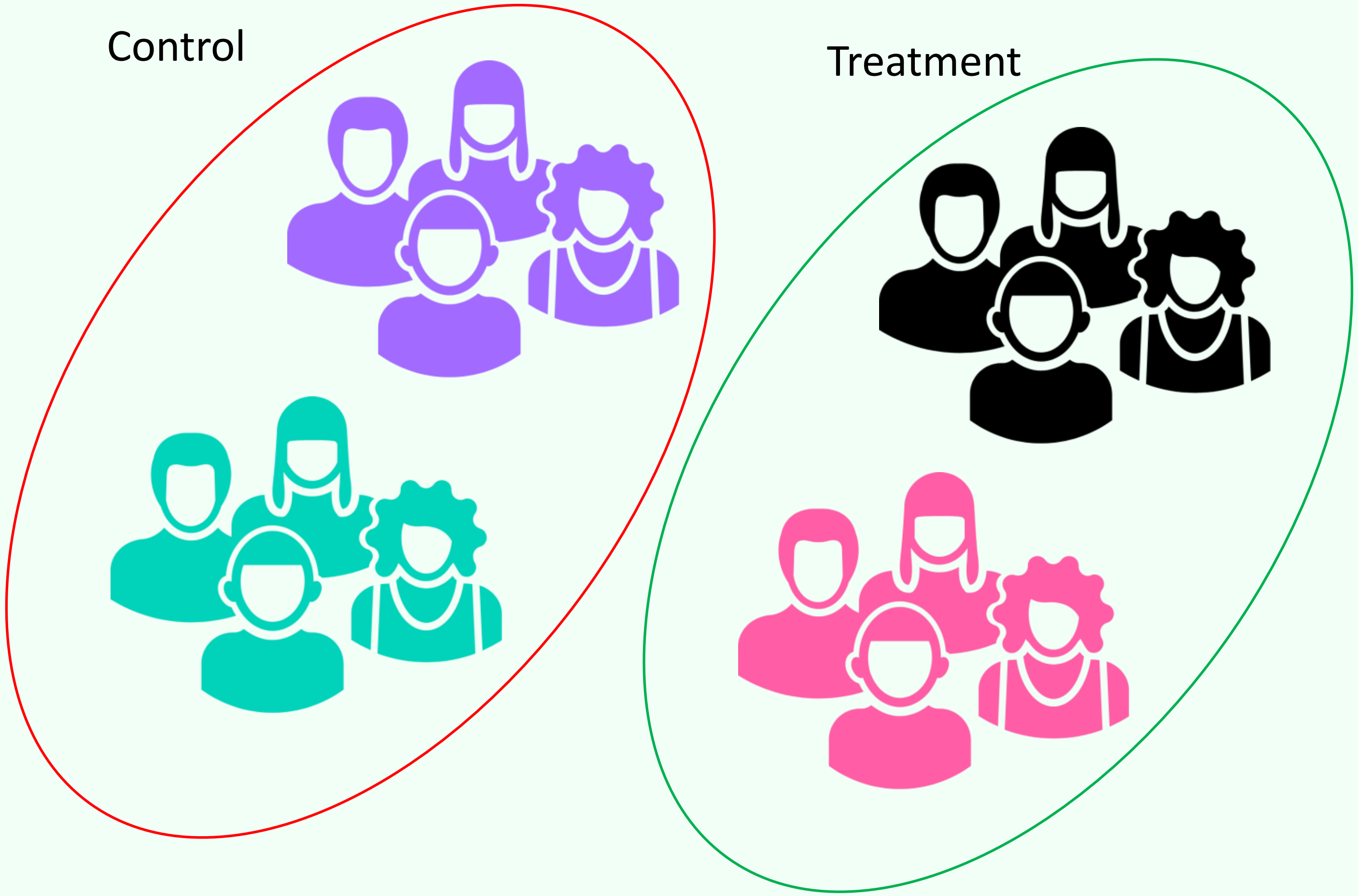
Source: Gupta and Kopper (2021)

Low intra-cluster correlation (ICC)



High intra-cluster correlation





ICC = degree of similarity in outcomes among units within pre-existing groups (e.g. children in schools)

➔ Risk: Picking people of a certain type

Power calculation formula

$$N = (t_{1-K} + t_{\frac{\alpha}{2}})^2 \frac{1}{P(1-P)} \times \frac{\sigma^2}{MDE^2}$$

$$MDE = (t_{1-K} + t_{\frac{\alpha}{2}}) \sqrt{\frac{1}{P(1-P)} \times \frac{\sigma^2}{N}}$$

N: Sample size, MDE: minimum detectable effect size, t_{1-K} and $t_{\frac{\alpha}{2}}$: critical values from a Student's t distribution, P: Power (0.8 or 0.9), σ^2: outcome variance.