Sampling and Sample Size

Matt Lowe

MIT

5 July 2017

Course Overview

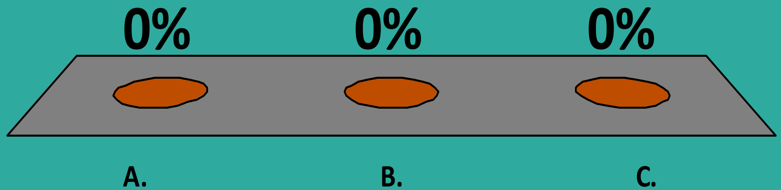
1. What is Evaluation?
2. Theory of Change
3. Outcome, Impact, and Indicators
4. Why Randomize?
5. How to Randomize
6. **Sampling and Sample Size**
7. Threats and Analysis
8. Research to Policy
9. Project from Start to Finish

Which of these most likely to describe estimates from 8 well implemented RCTs?

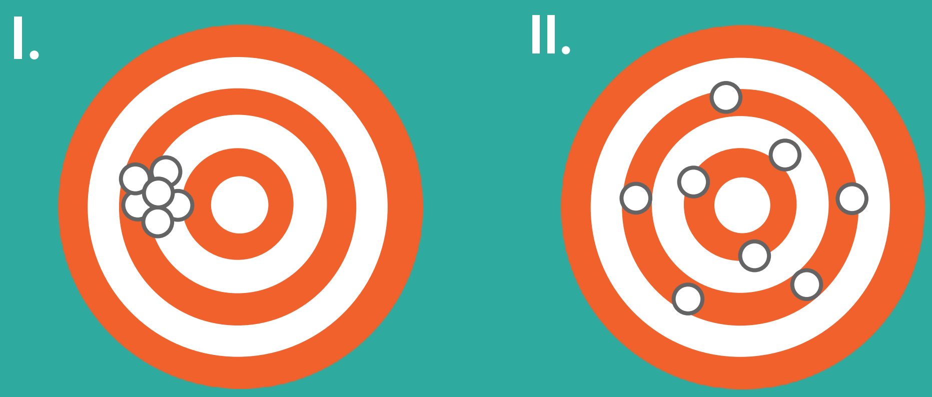
I.

II.

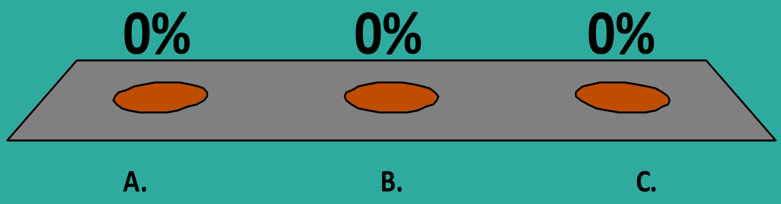
A. I
B. II
C. Neither



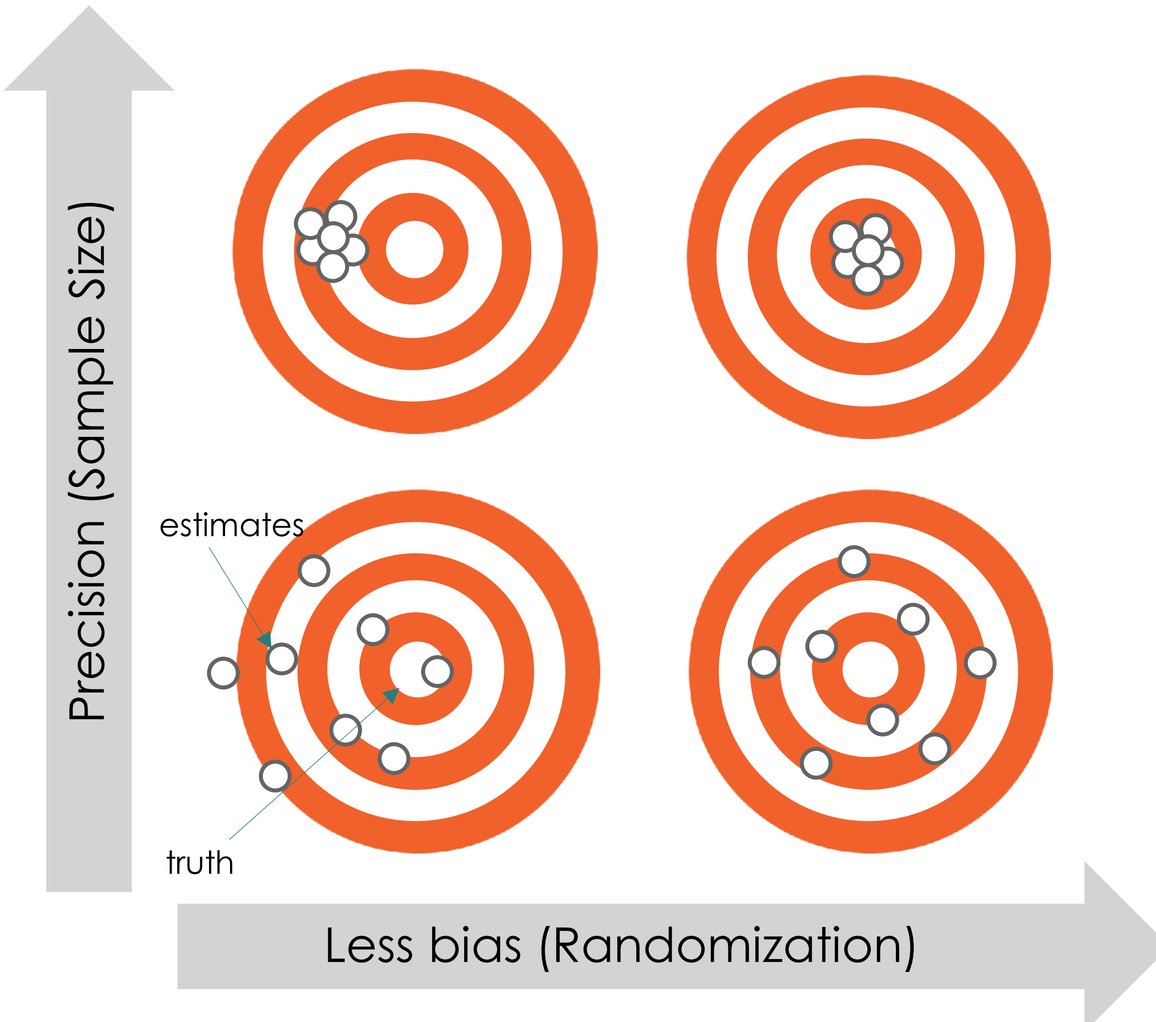
Which is the best description of II?



A. Imprecise estimate
B. Biased estimate
C. Imprecise but unbiased



Bias and precision



Outline

- Introduction
- Hypothesis testing
- What influences power?
- Power in clustered designs
- Calculating power in practice

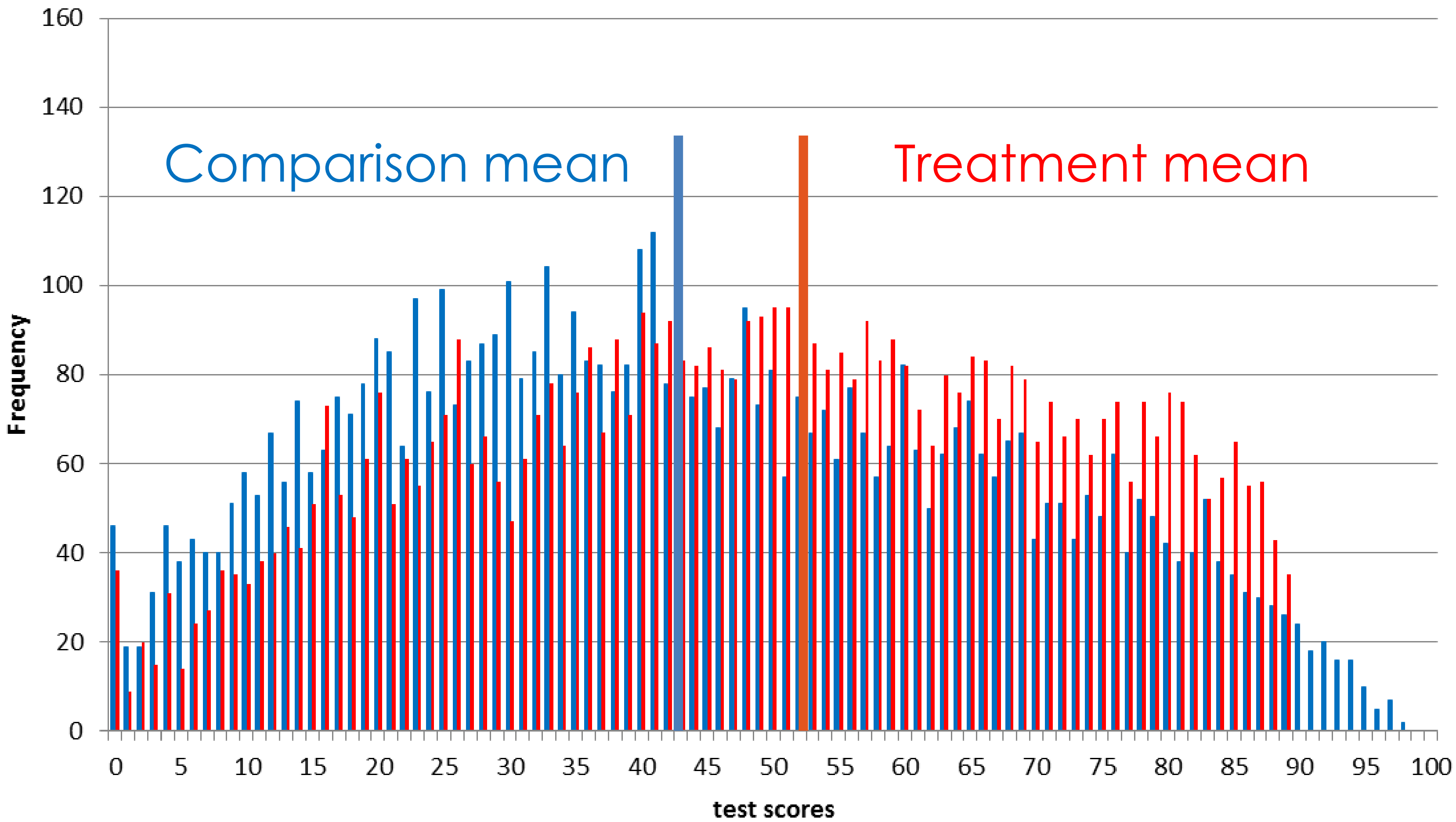
Outline

- **Introduction**
- Hypothesis testing
- What influences power?
- Power in clustered designs
- Calculating power in practice

Introduction

We evaluate bringing tutors into schools

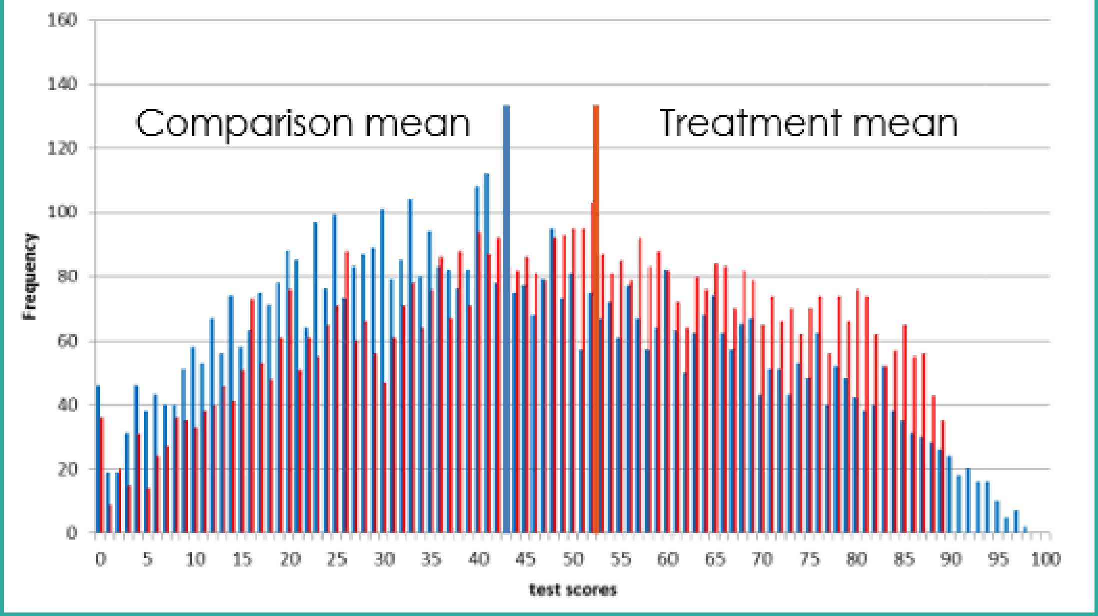
Post-test: control & treatment



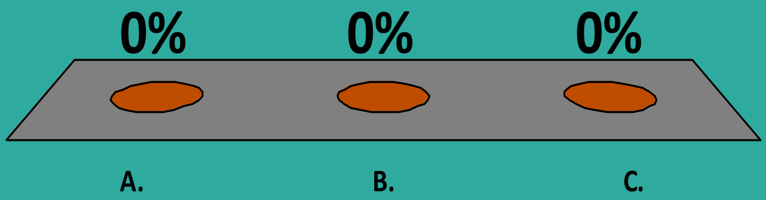
The mean of treatment is 6ppt higher than mean of control

Is this impact statistically significant?

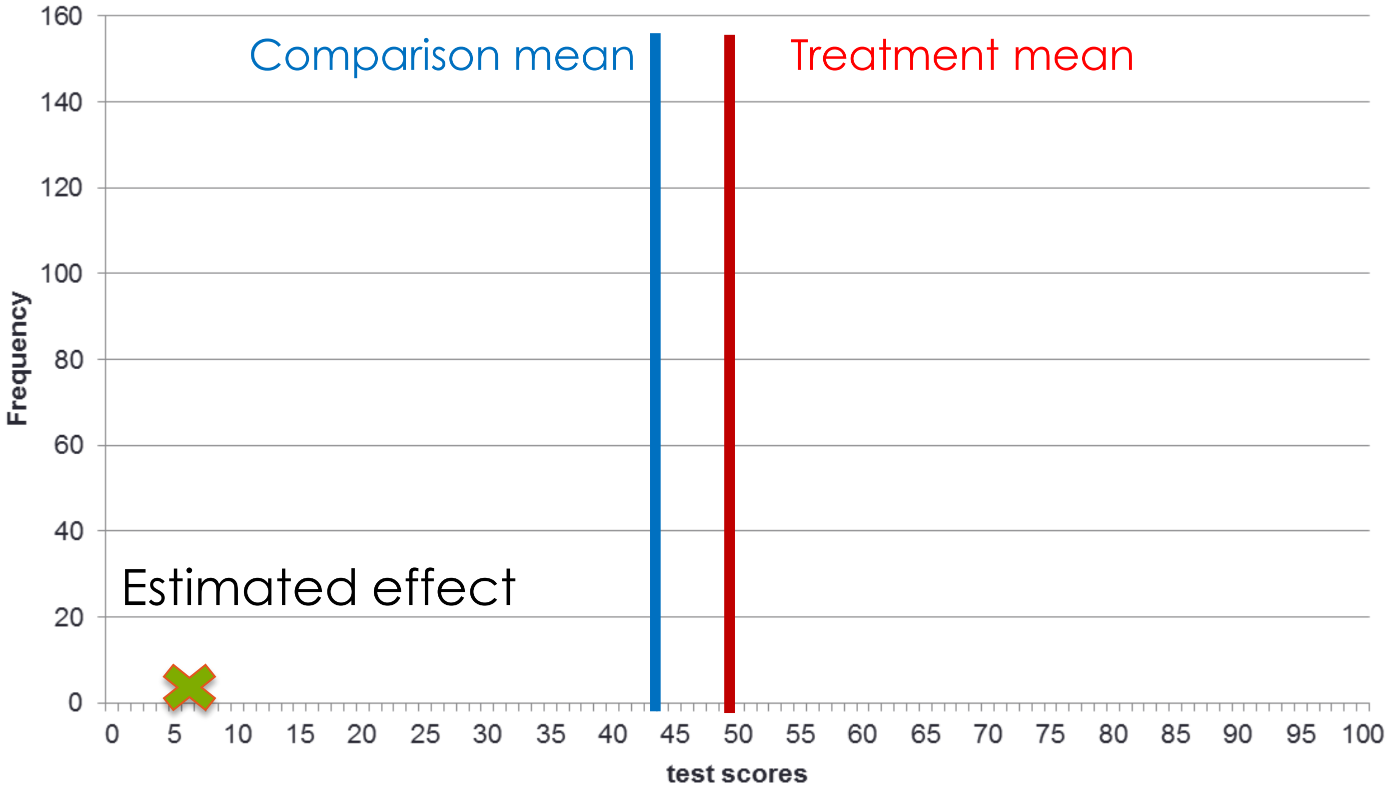
Average Difference = 6 points



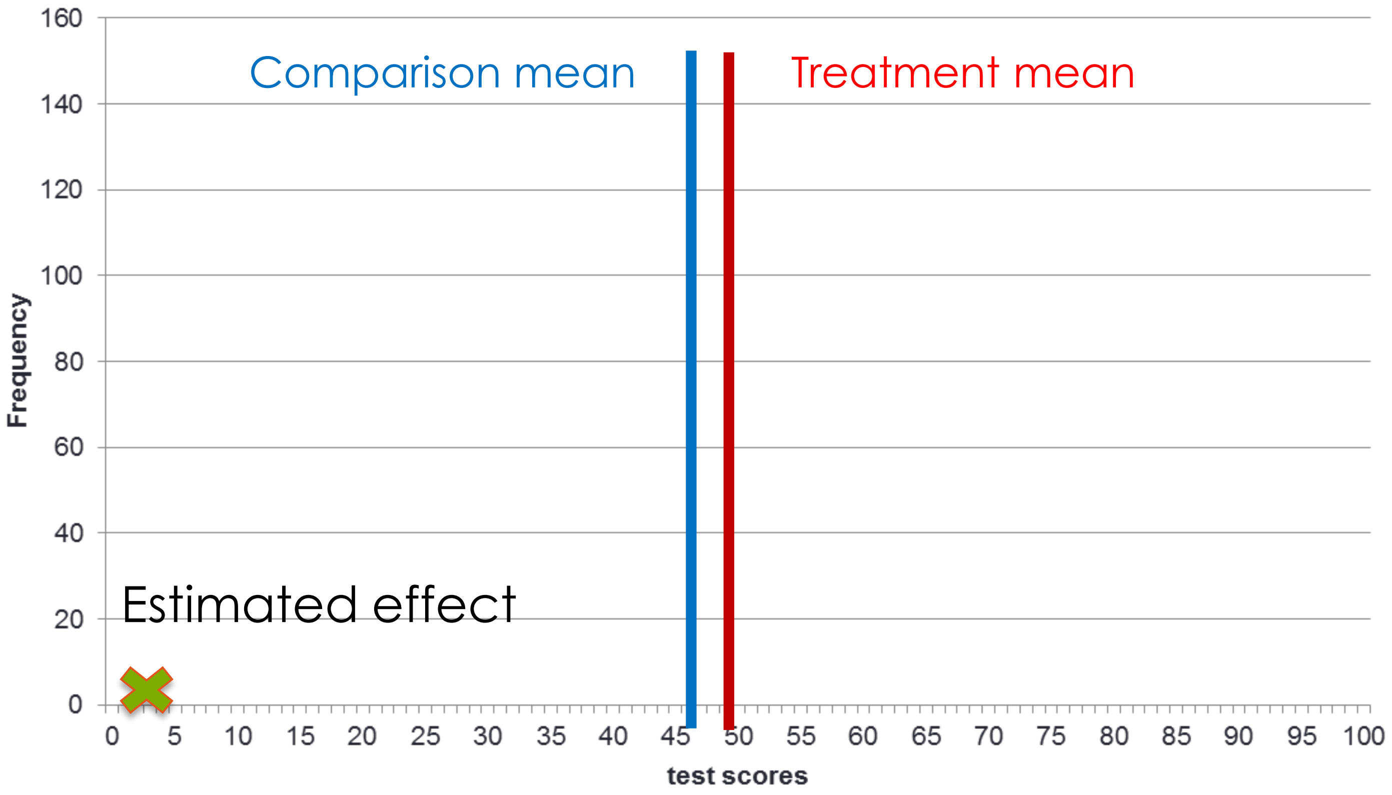
A. Yes
B. No
C. We cant tell



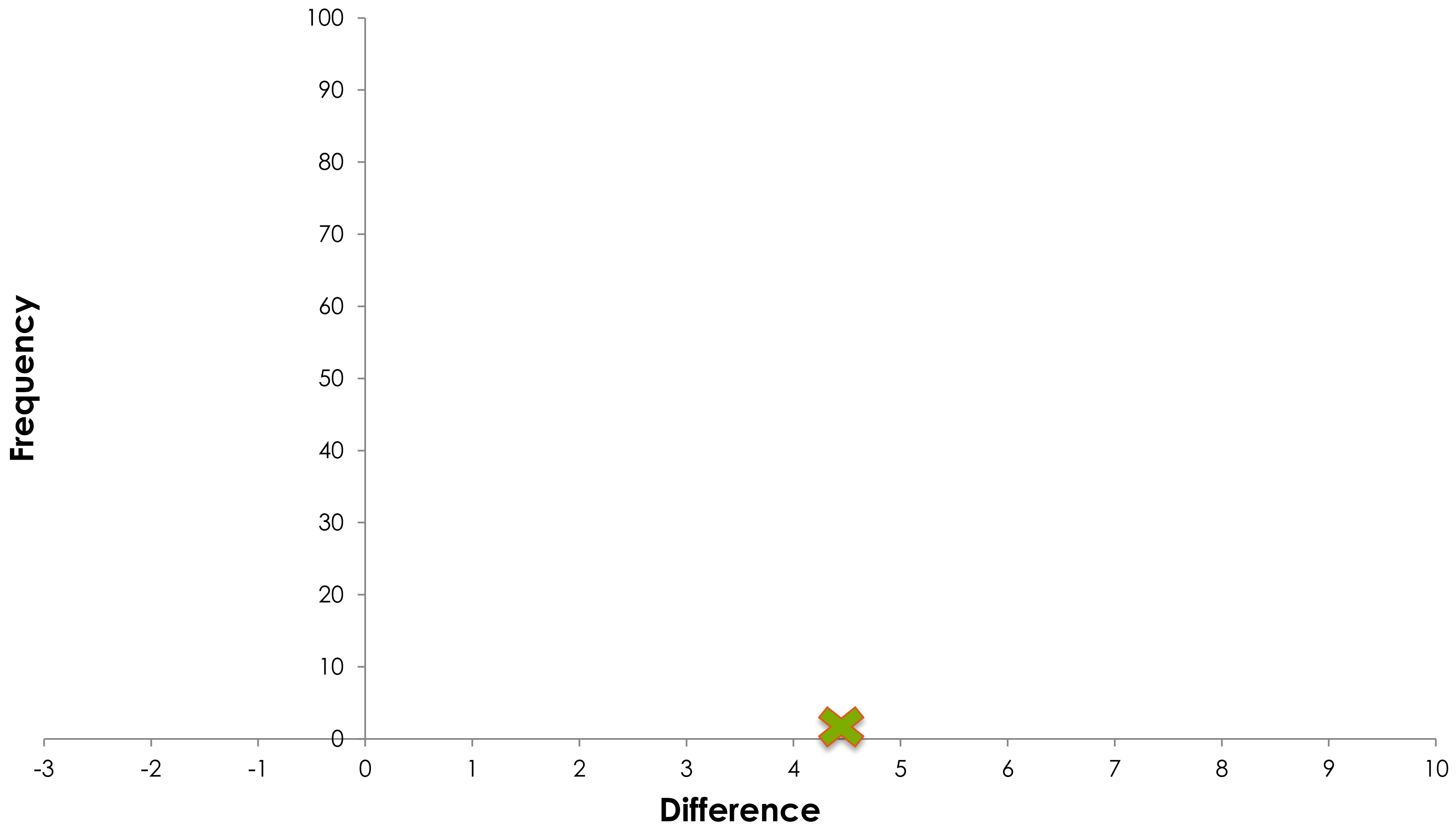
Difference between the sample means



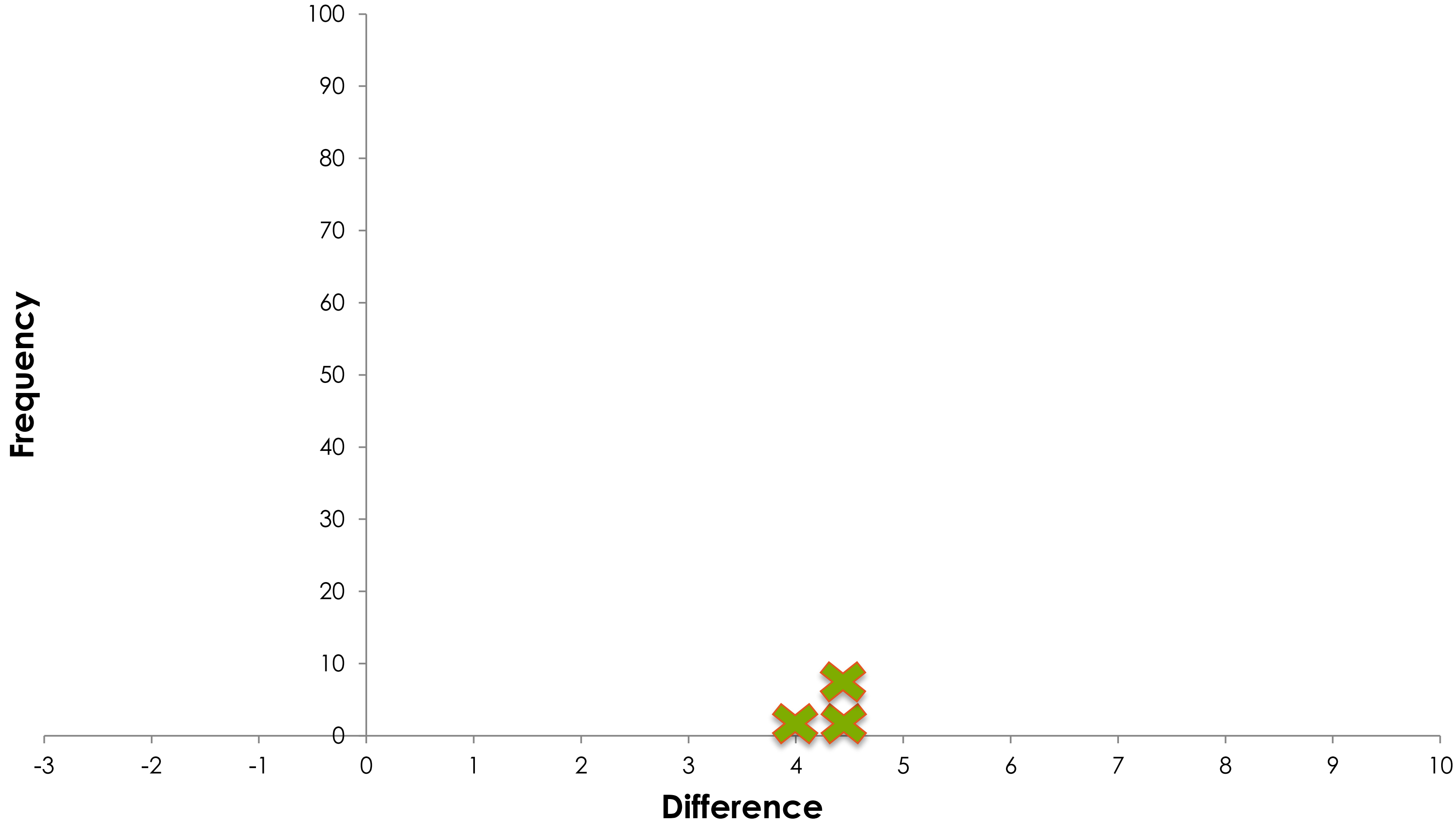
What if we ran a second experiment?



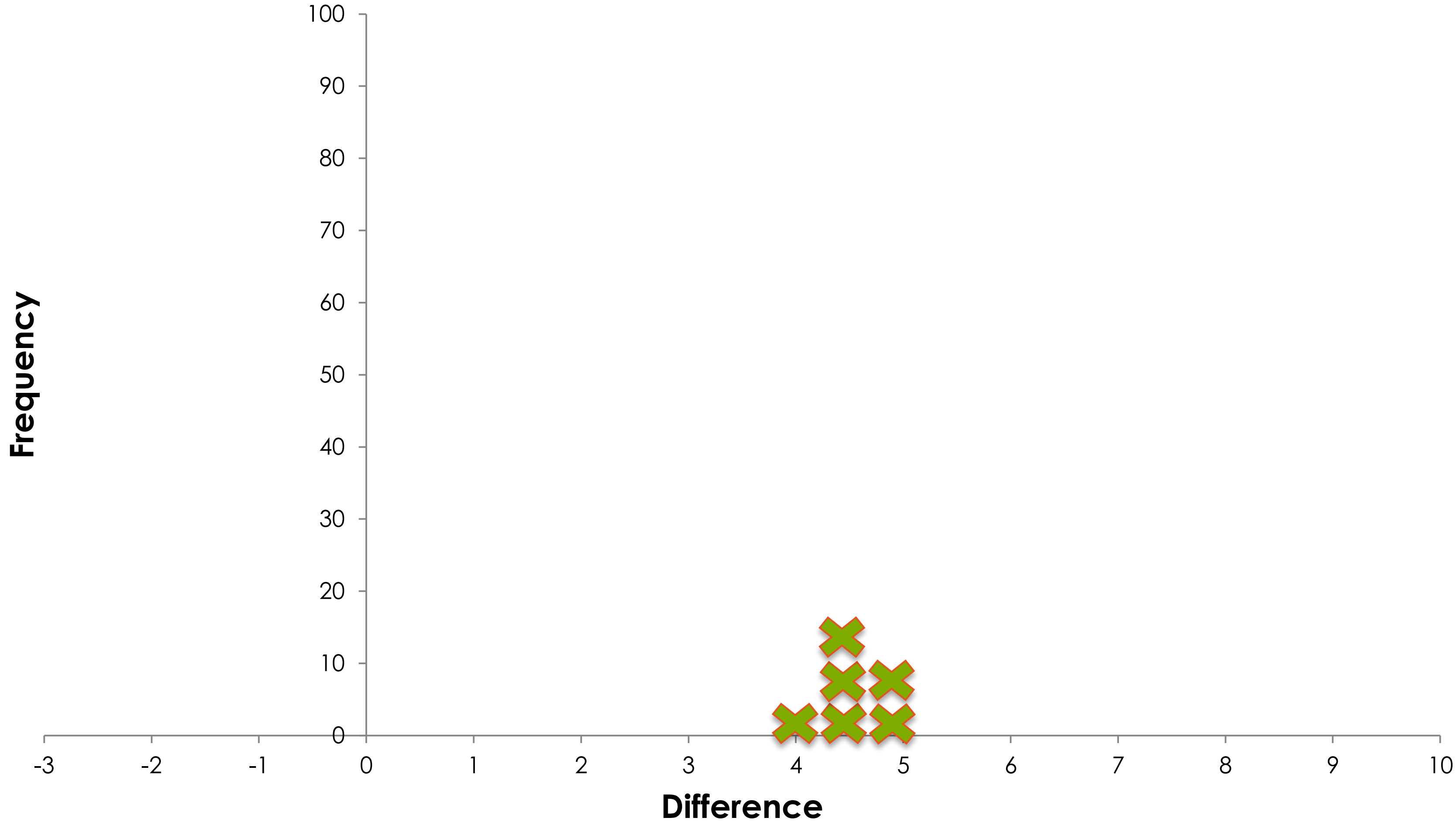
Many experiments: a distribution of estimates



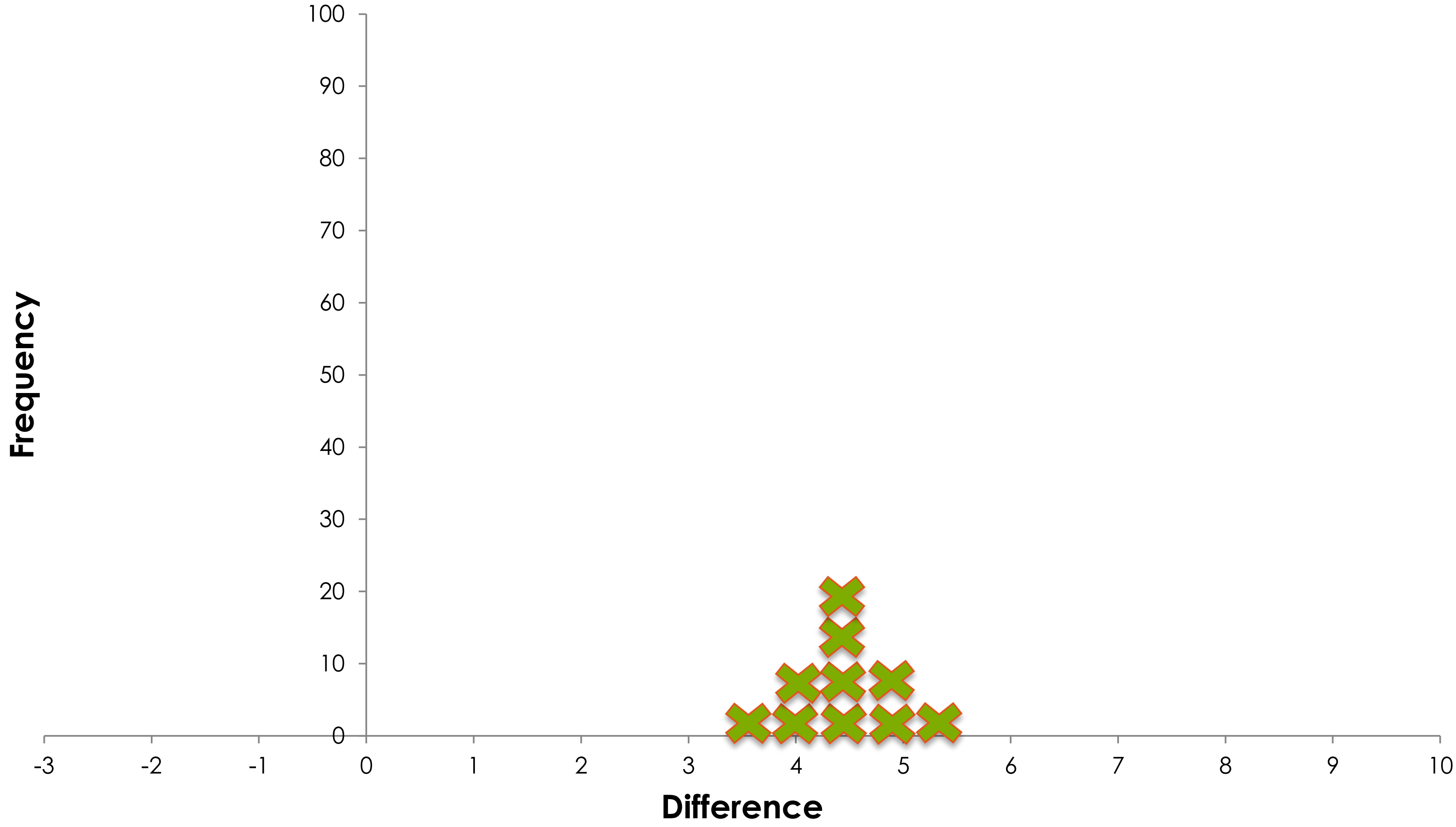
Many experiments: a distribution of estimates



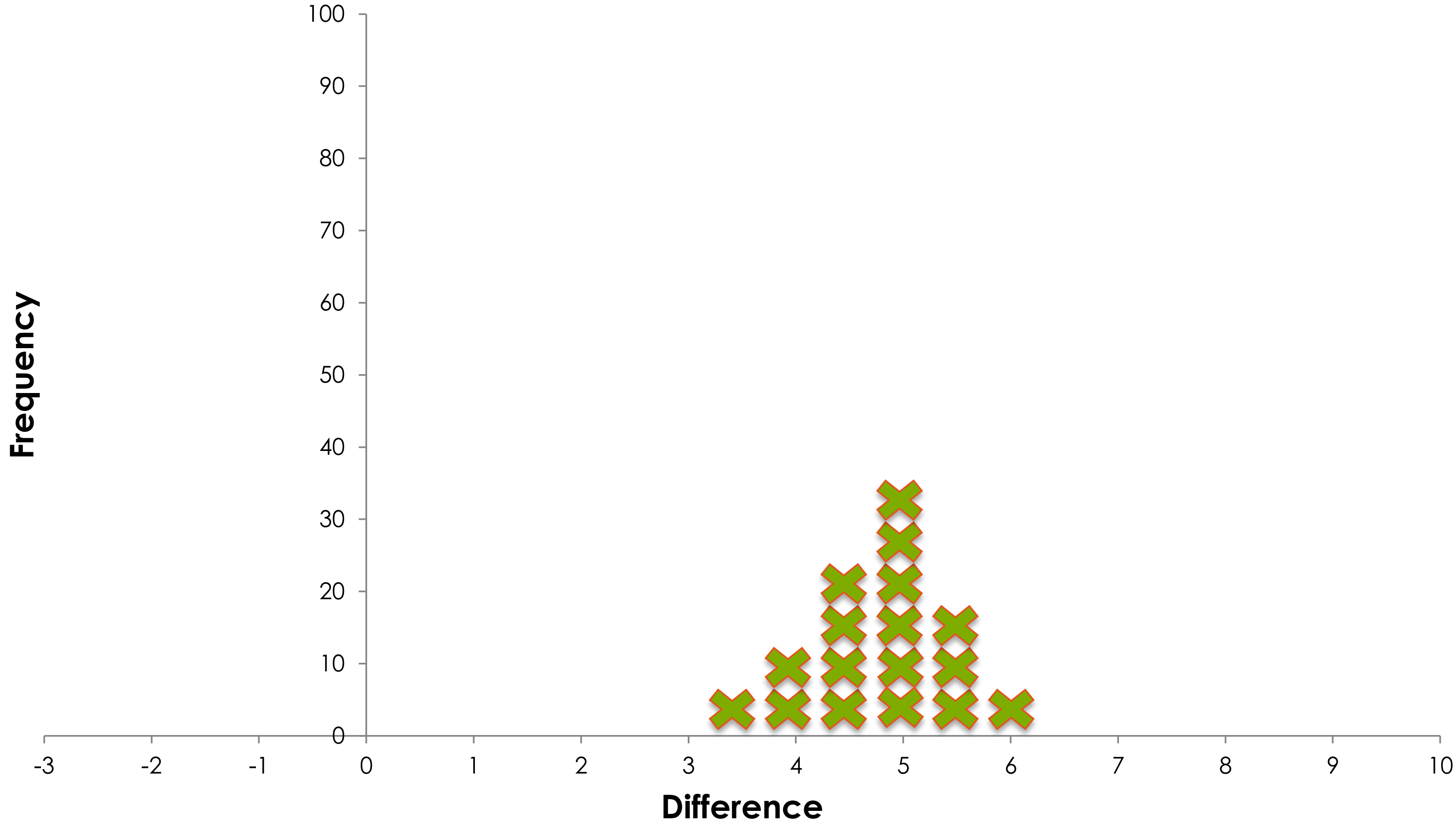
Many experiments: a distribution of estimates



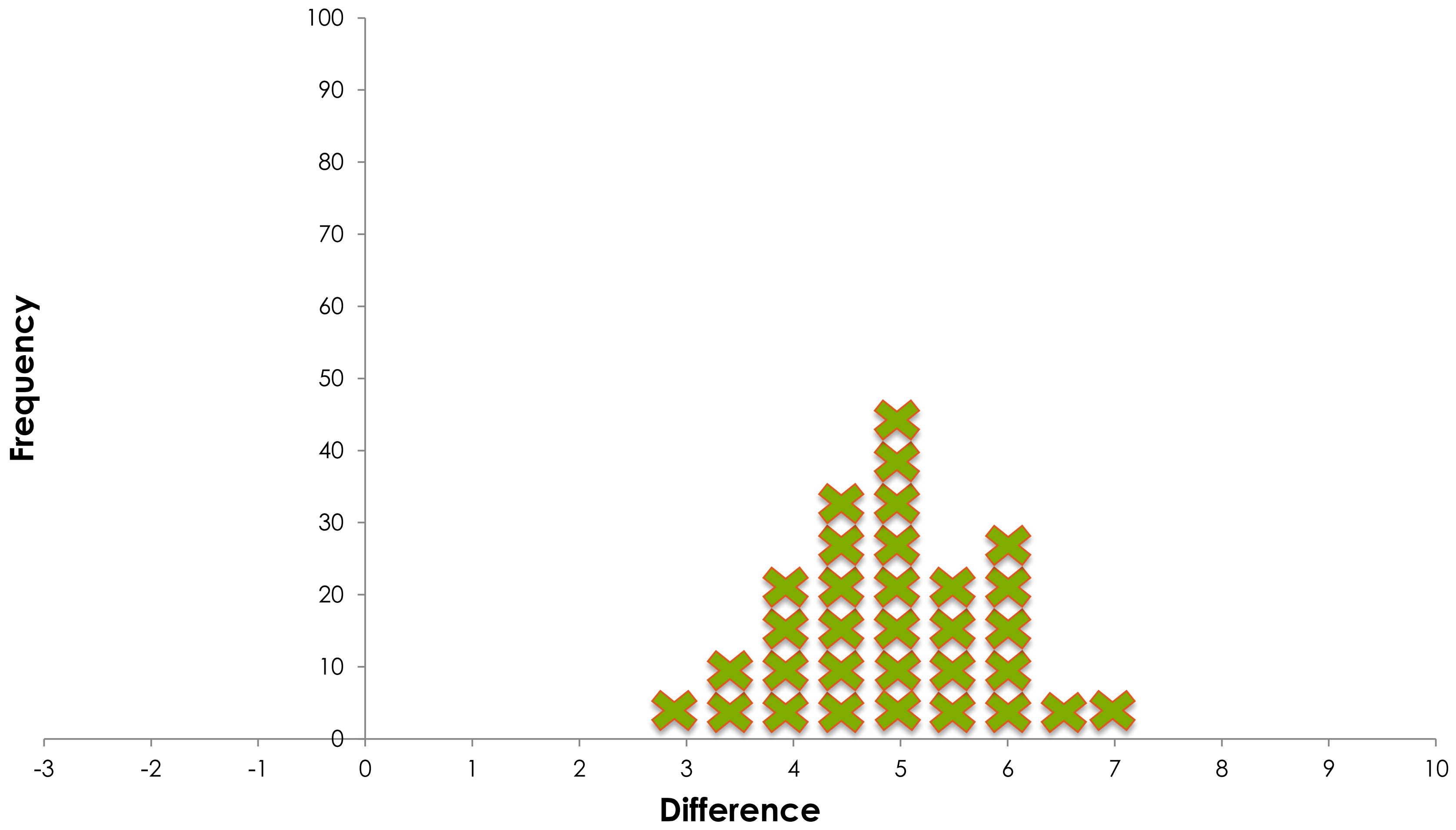
Many experiments: a distribution of estimates



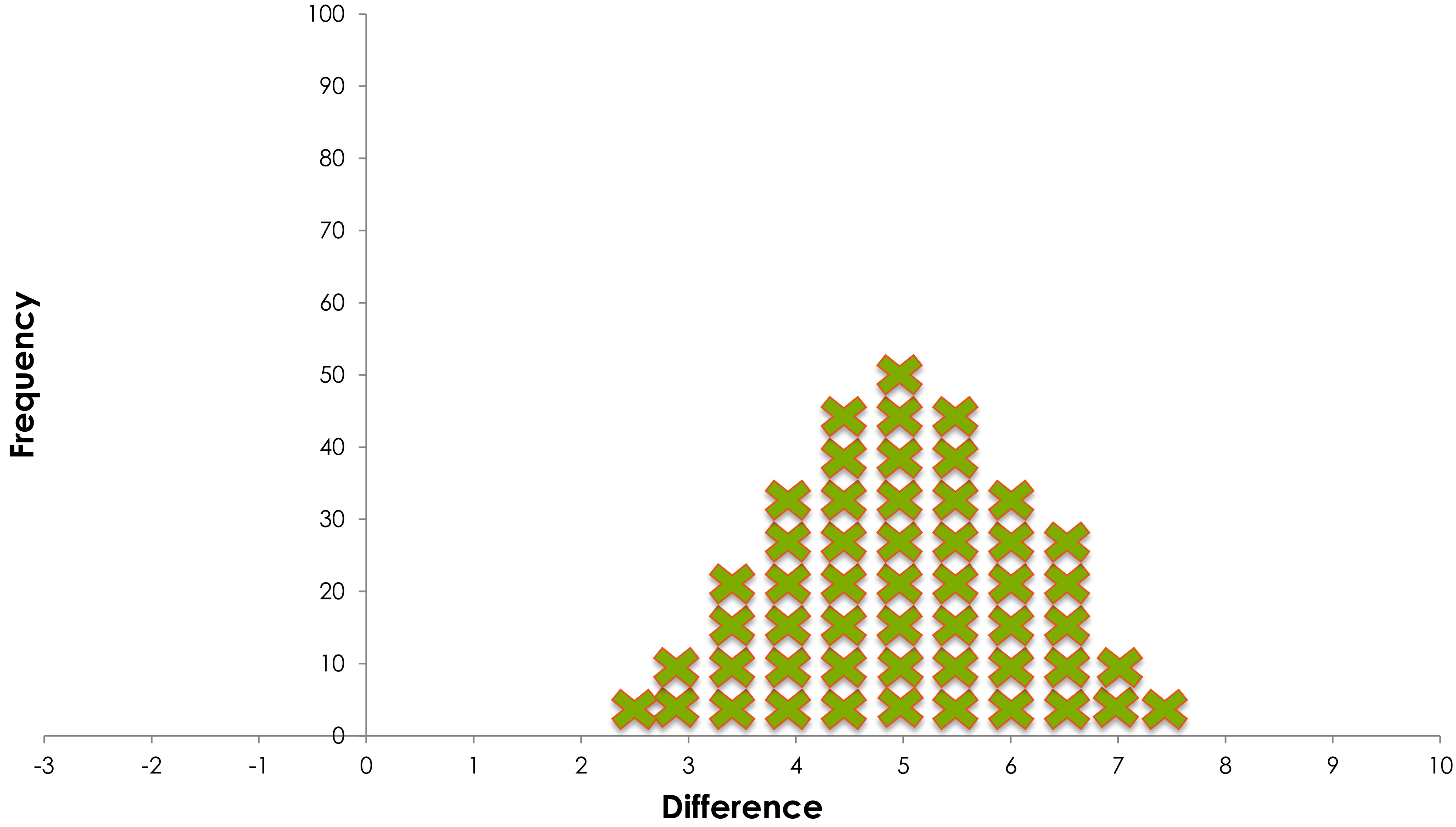
Many experiments: a distribution of estimates



Many experiments: a distribution of estimates



Many experiments: a distribution of estimates

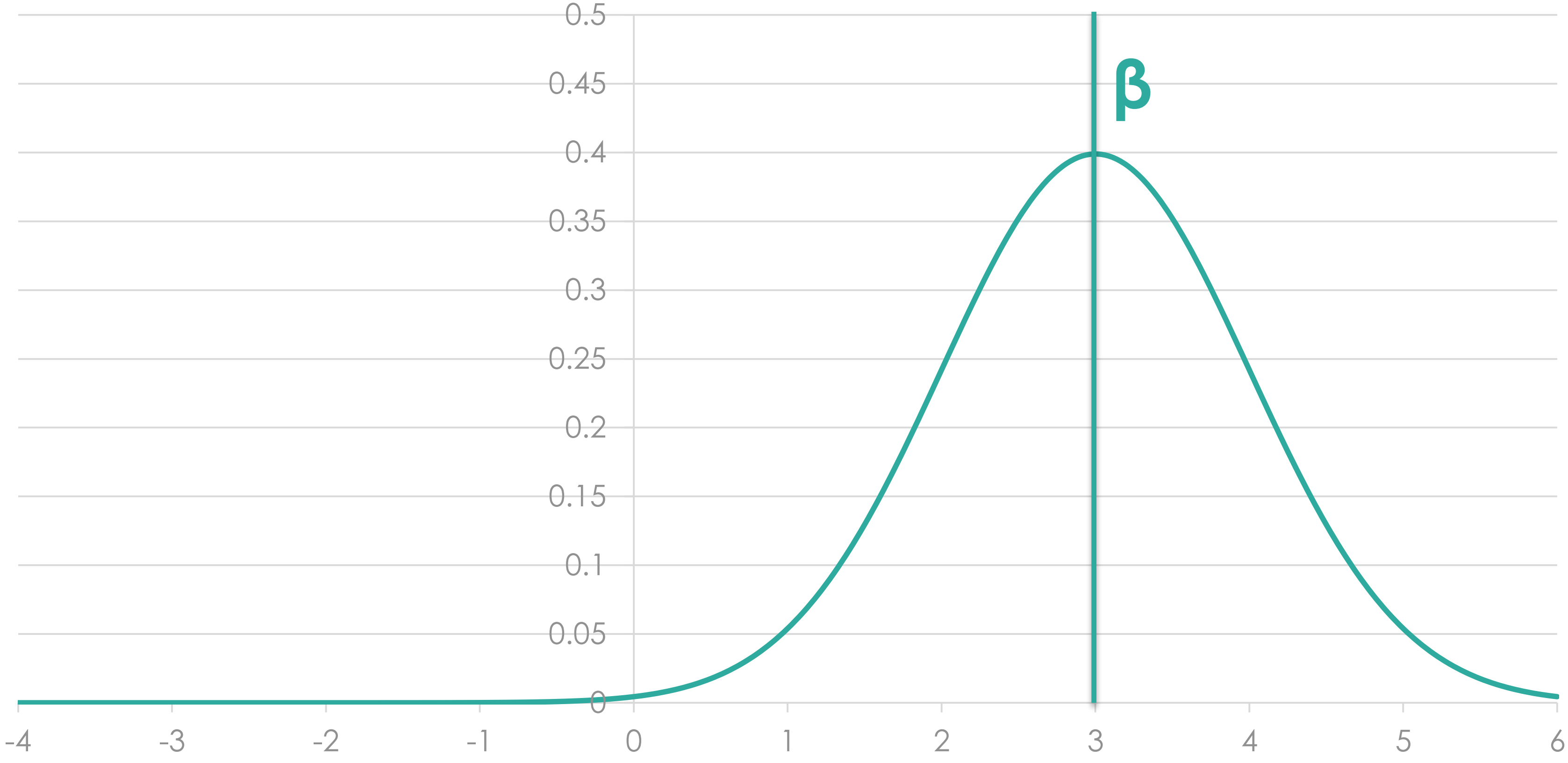


Outline

- Introduction
- **Hypothesis testing**
- What influences power?
- Power in clustered designs
- Calculating power in practice

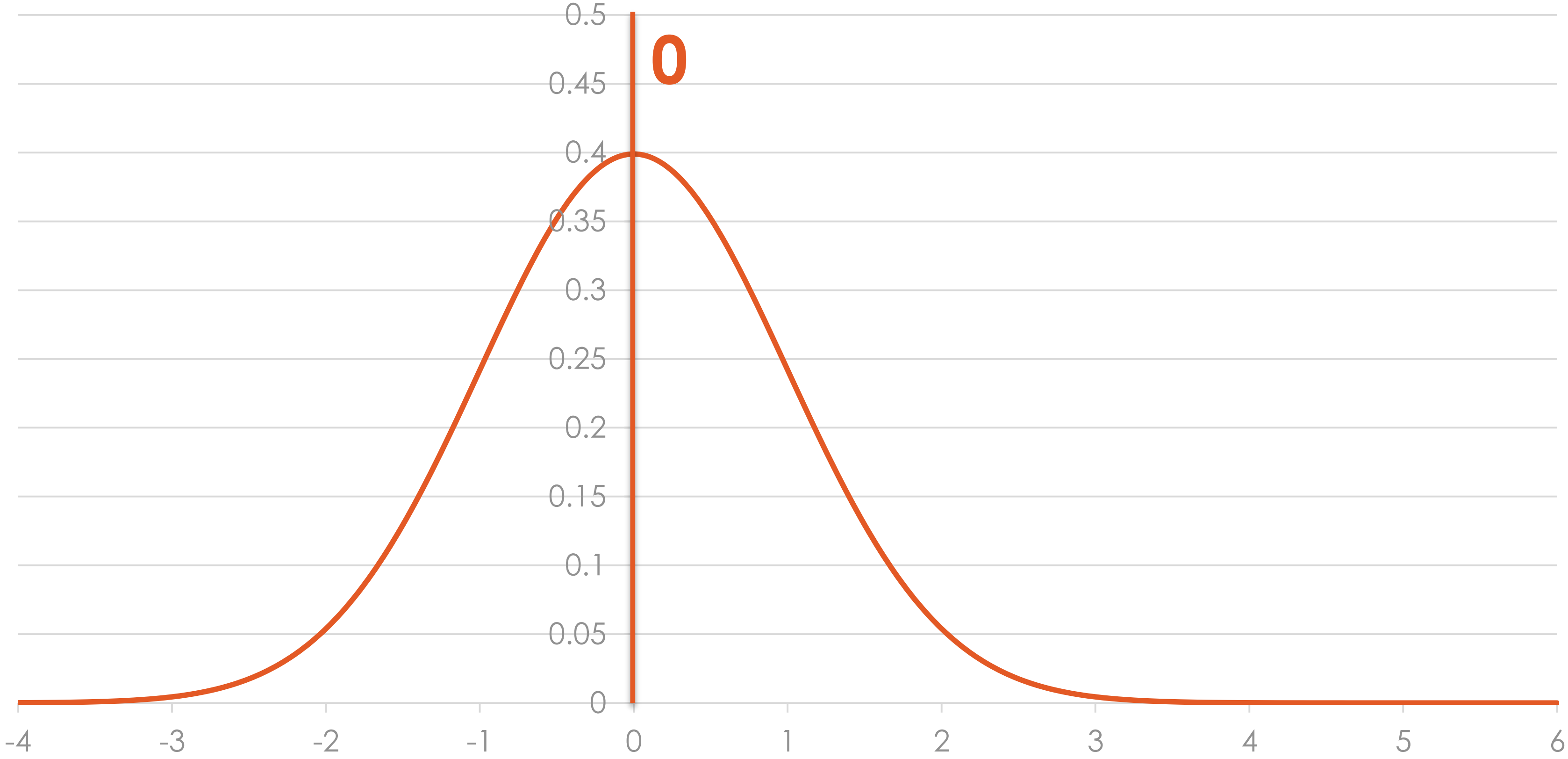
Hypothesis testing

Distribution of estimates if true effect=β



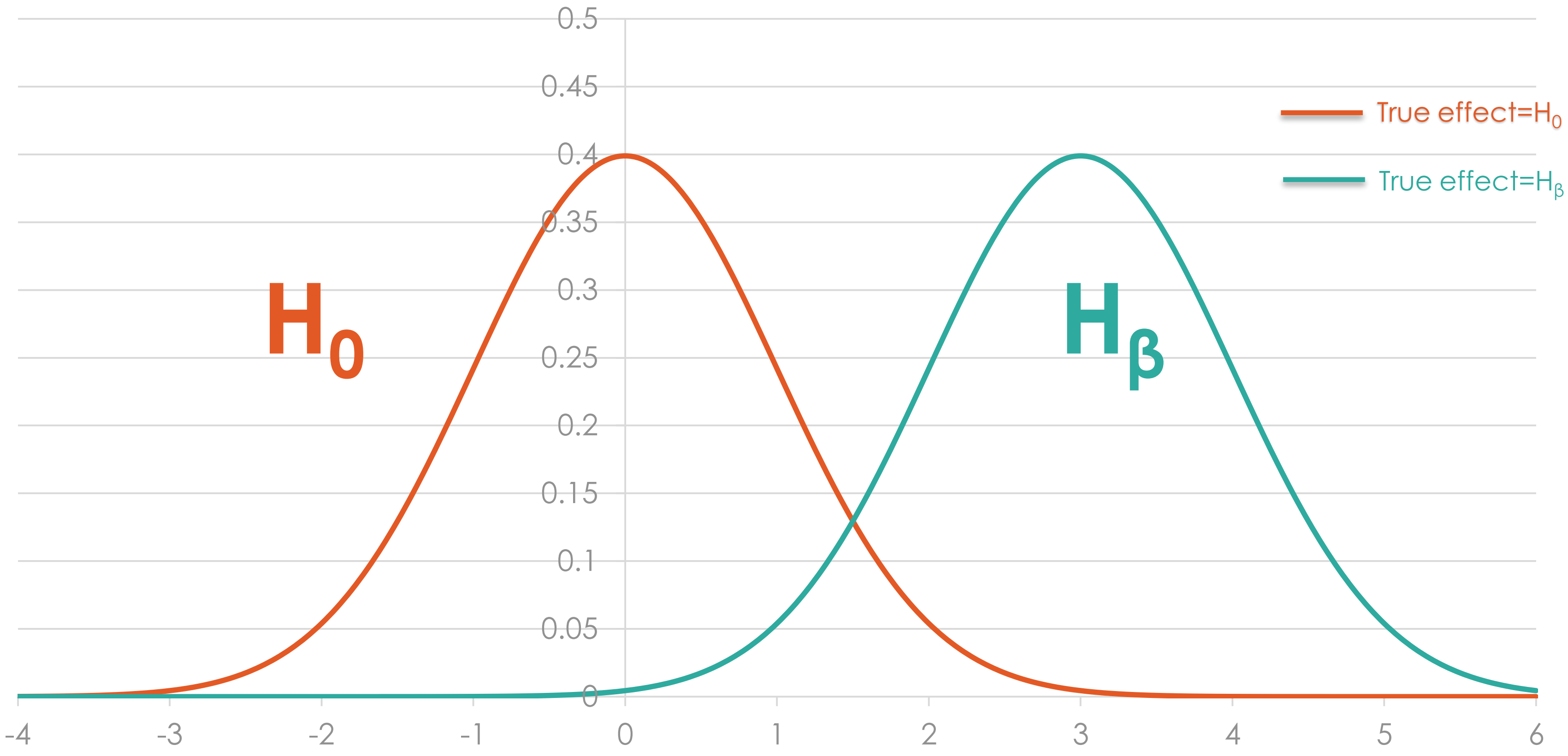
Estimated effects normally distributed around effect β

Distribution of estimates if true effect=0



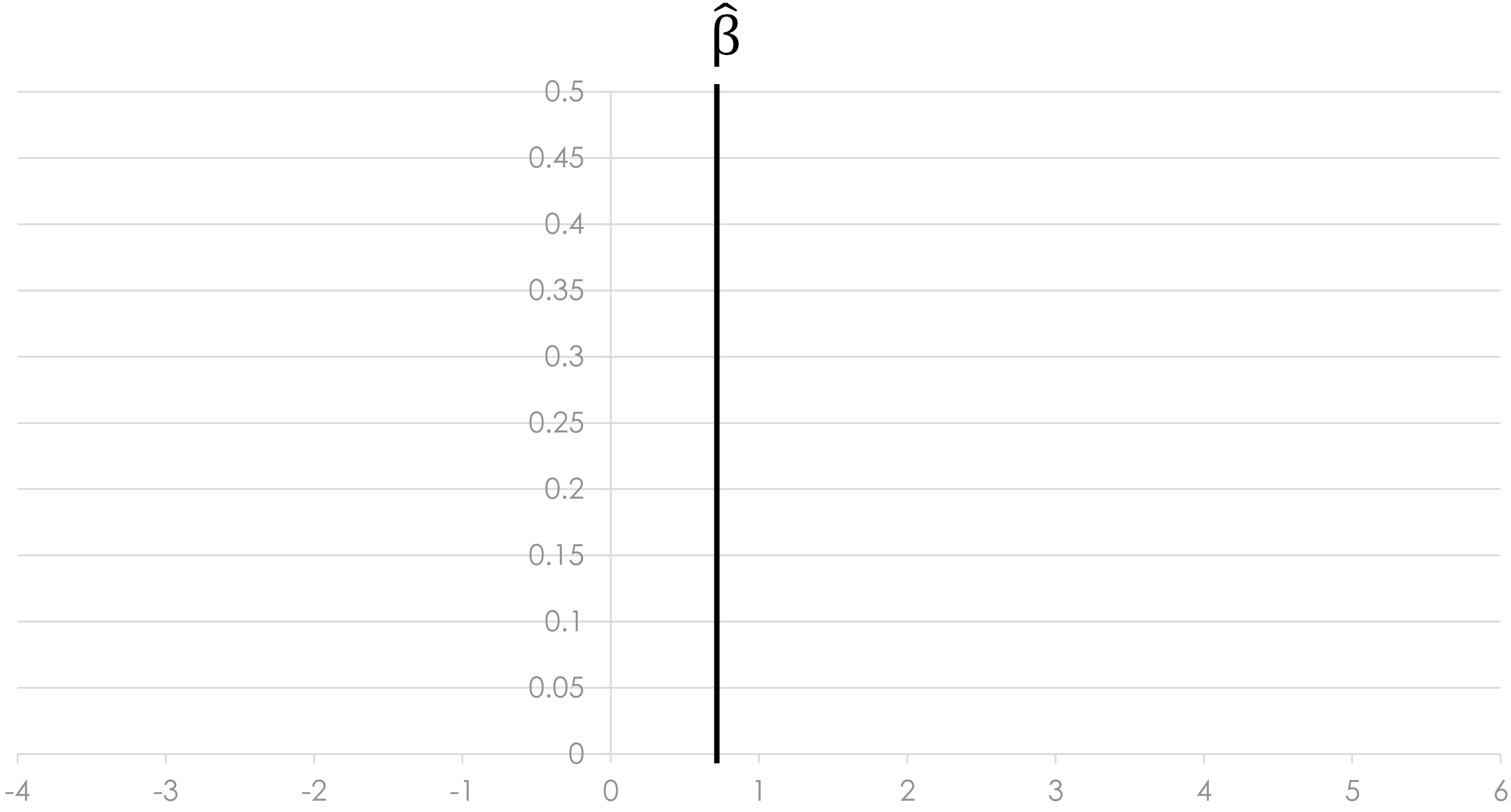
Estimated effects normally distributed around effect 0

Two distributions under two hypotheses



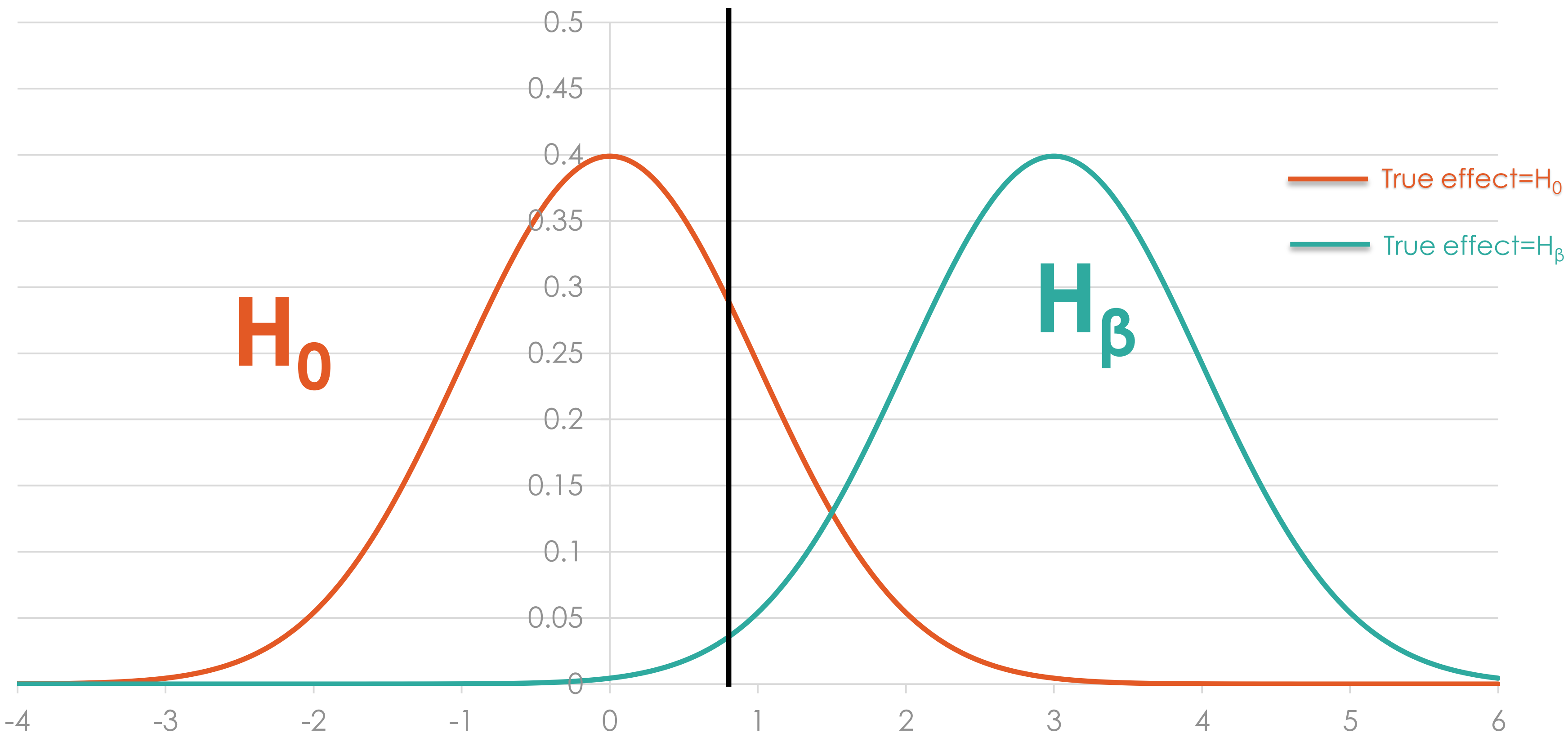
Probability of getting given estimate under two hypotheses

If we run one study, see one estimate $\hat{\beta}$



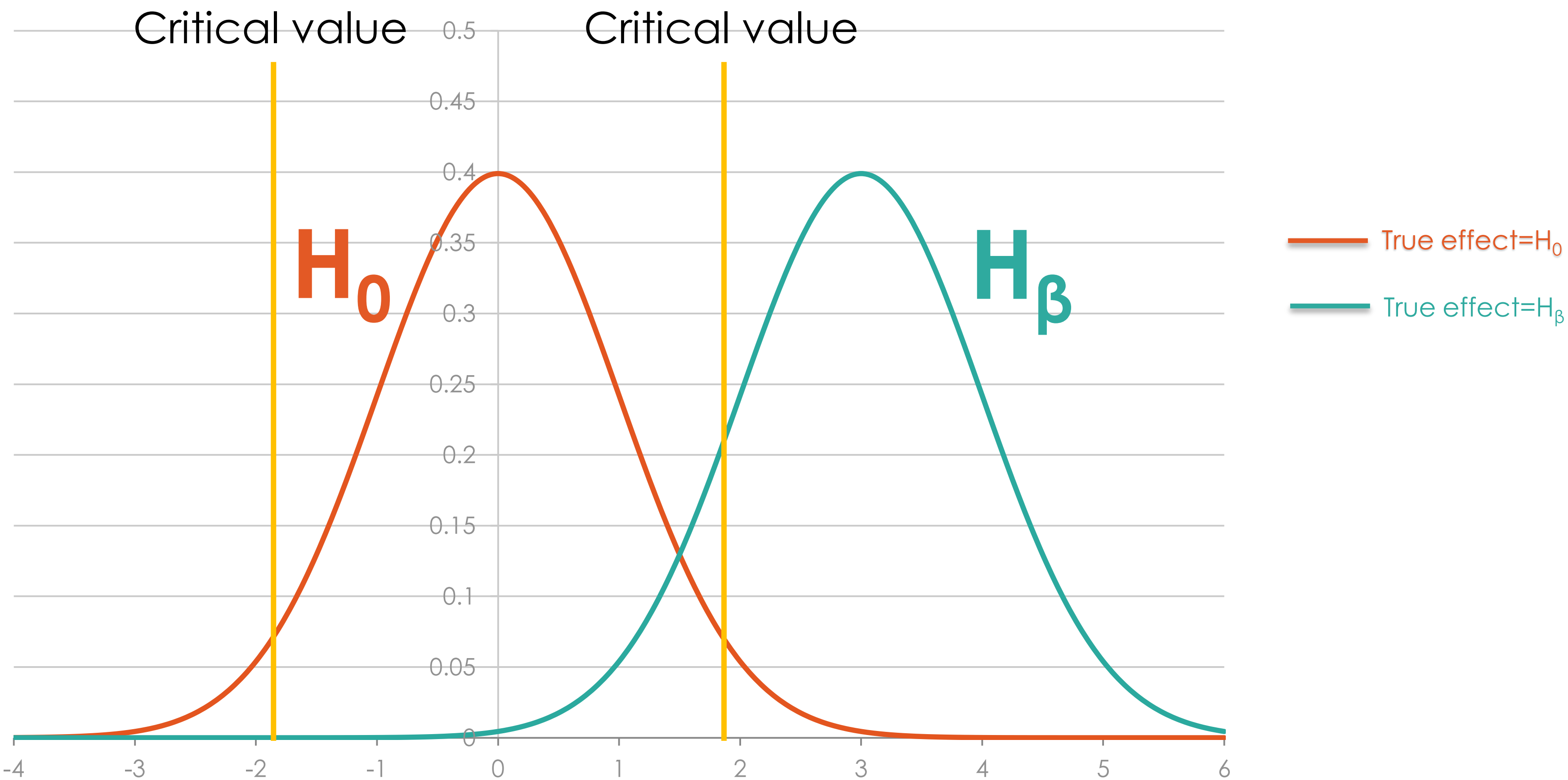
How do we know if our estimated effect is significant?

Did our estimate come from H_β or H_0?



Can we rule out that the estimate comes from H_0?

Impose significance level of 5%



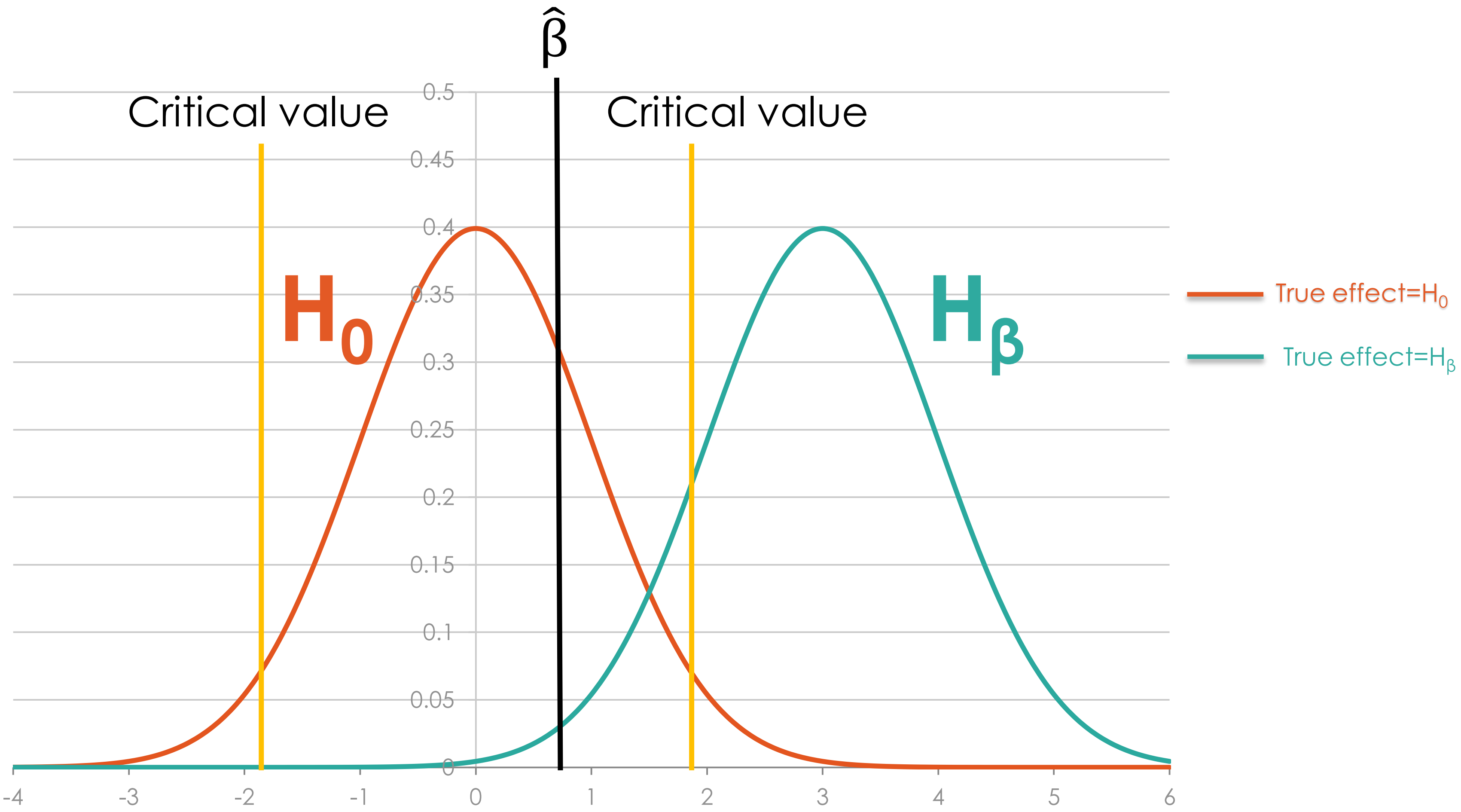
Anything between lines cannot be distinguished from 0

Critical value

- Definition: The estimated effect size that exactly corresponds to the significance level.
- If testing whether:
  - The effect is significantly different to 0
  - Significant at 5% level

  Then, critical value is the level of the estimate where exactly 2.5% of area under the curve lies to the right/left (the two areas sum to 5%)

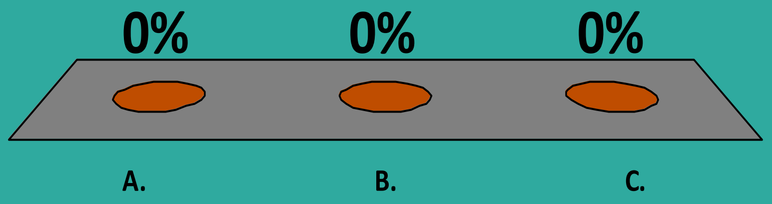
Is $\hat{\beta}$ significantly different from 0, at 5%?



How do we know?

Is $\hat{\beta}$ significantly different from 0, at 5%?

A. Yes
B. No
C. Can not tell without more information



Hypothesis Testing

- In criminal law, most institutions follow the rule: "innocent until proven guilty"
- The presumption is that the accused is innocent and the burden is on the prosecutor to show guilt
  - The jury or judge starts with the "null hypothesis" that the accused person is innocent
  - The prosecutor has a hypothesis that the accused person is guilty
  - Don't want innocent to go to prison: "reject the null" only when strong evidence

Hypothesis Testing

- Usually in program evaluation, instead of "presumption of innocence," the rule is: "presumption of zero"
- The "Null hypothesis" (H_0) is then that there was no impact of the program
- The burden of proof is on showing there was an impact
- Note, there are exceptions:
  - e.g. the null for a cash plus training program might be that it's the same as the training on its own
  - Null hypothesis could be a prior belief

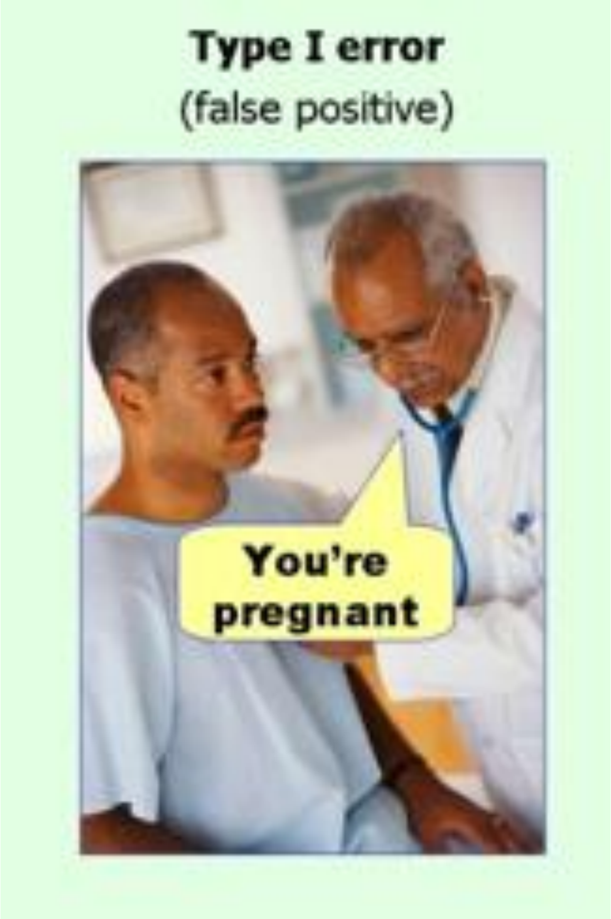
Hypothesis Testing: Conclusions

- If it is very unlikely (less than a 5% probability) that the difference is solely due to chance:
  - We "reject our null hypothesis"
- We may now say:

  "our program has a statistically significant impact"
- What is statistically significant and what is most likely are different concepts
  - there may be cases where it's more likely that the program worked than that it didn't, but we still say there is no statistically significant impact

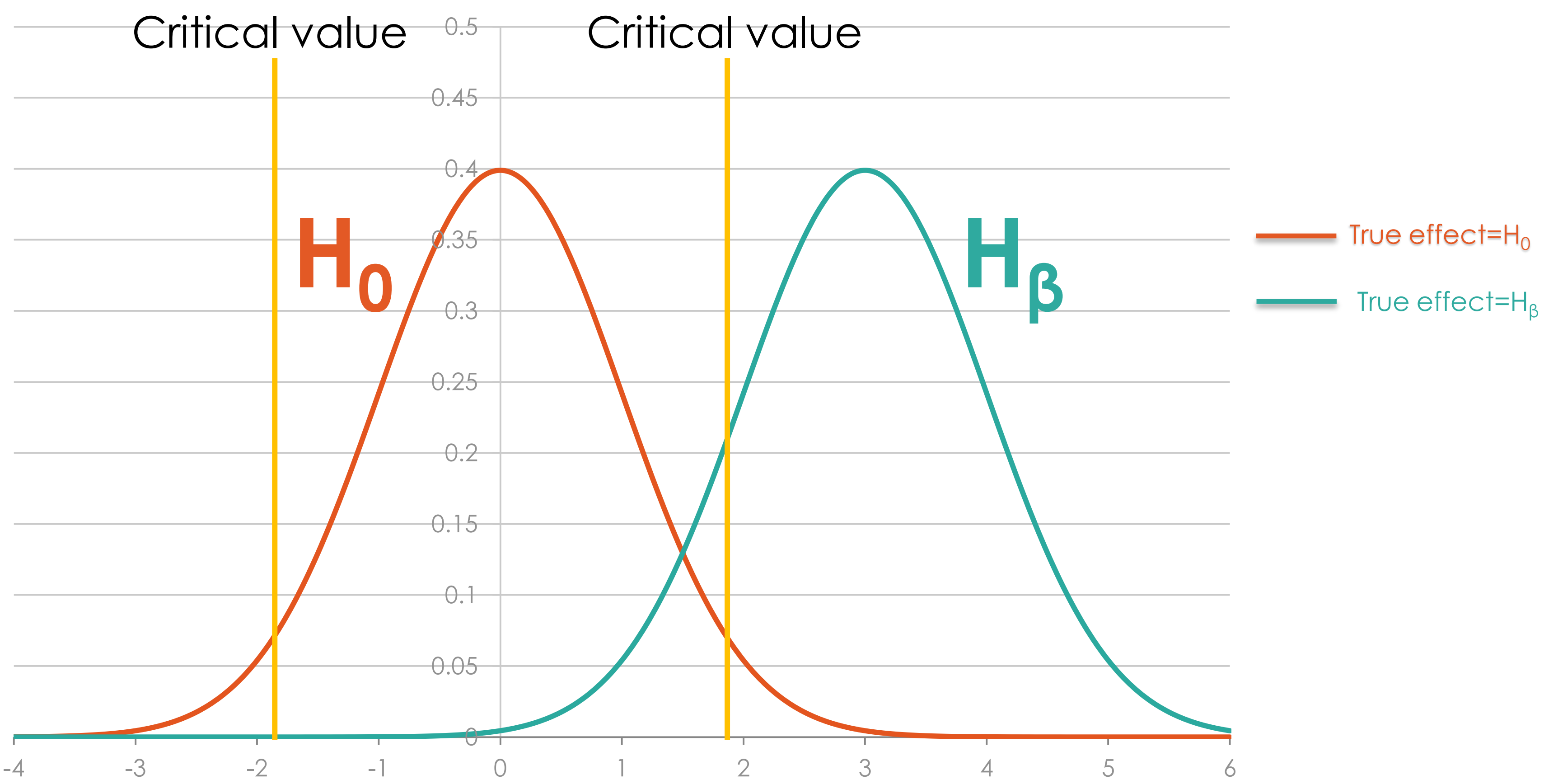
Significance and Type I errors

- Traditionally significance level is set at 5%
- This means allowing a 5% chance of experiencing Type I errors
- 5% of time we will say program had impact when in fact it didn't



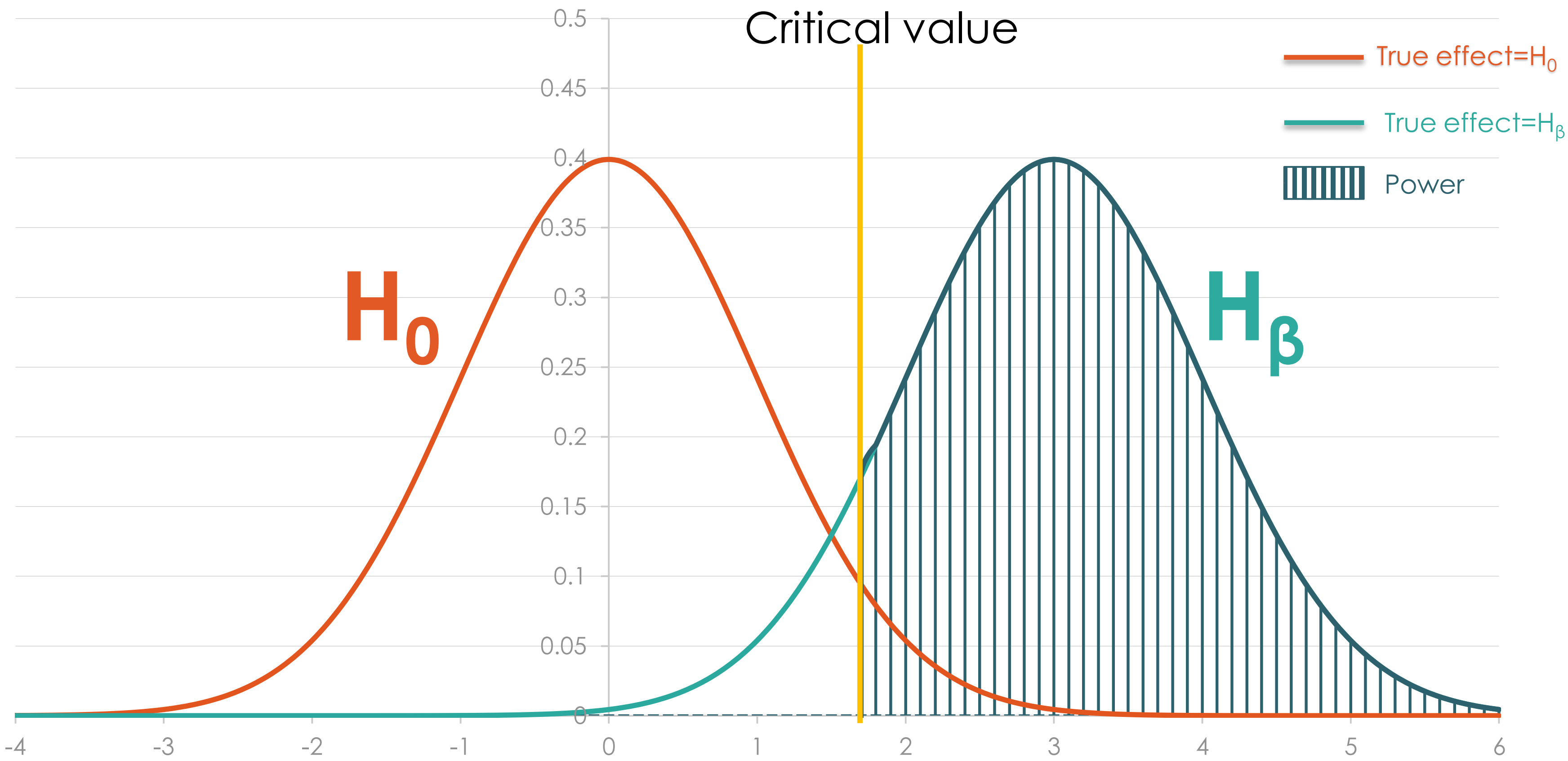
Source: Effect Size FAQs blog

If true effect was β



How often would we get an estimated effect we can distinguish from zero?

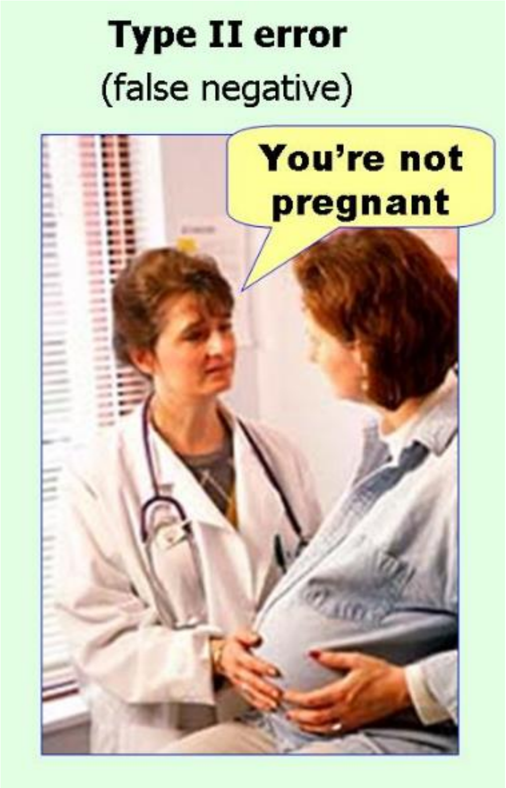
How often would we reject null, if Hβ true?



Shaded area shows power = % of time we would find Hβ different from 0 if true effect was β

The power to avoid Type II errors

- Statistical power is the probability that, if the true effect is of a given size, our proposed experiment will be able to distinguish the *estimated* effect from zero
- Power is the probability of avoiding Type II errors
- Traditionally, we aim for 80% power (some aim for 90%)
- Low power means we may not find a significant effect even though an effect exists



Source: Effect Size FAQs blog

Four results from hypothesis testing

Underlying truth

Effect

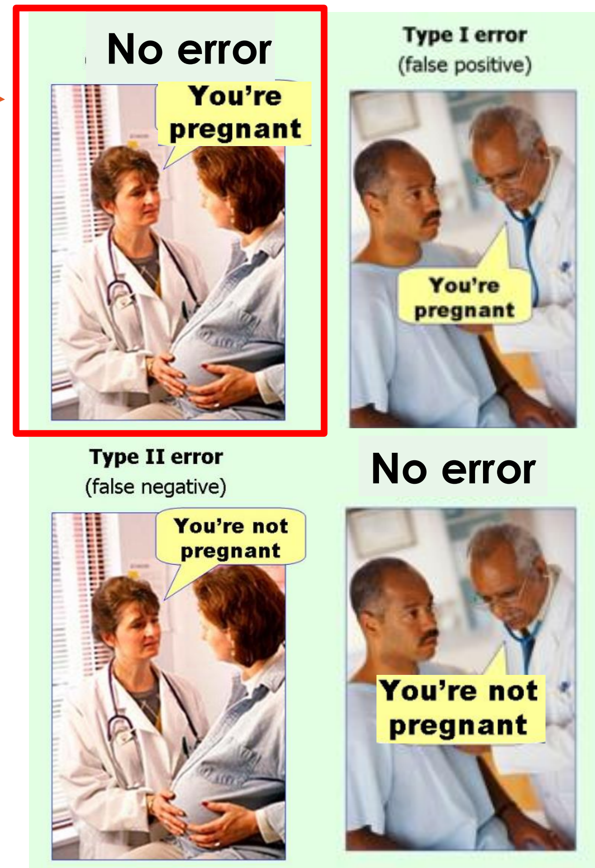
No effect

Power: when is effect prob, find significance

Significant

Statistical test

Not significant



Source: Effect Size FAQs blog

Four results from hypothesis testing

Statistical Test:	*The underlying truth:* TREATMENT EFFECT (H_0 false)	NO TREATMENT EFFECT (H_0 true)
SIGNIFICANT (Reject H0)	**True Positive** Probability = 1-κ	**False Positive** Probability = α Type I error
NOT SIGNIFICANT (Fail to reject H_0)	**False Zero** Probability = κ Type II error	**True Zero** Probability = (1-α)

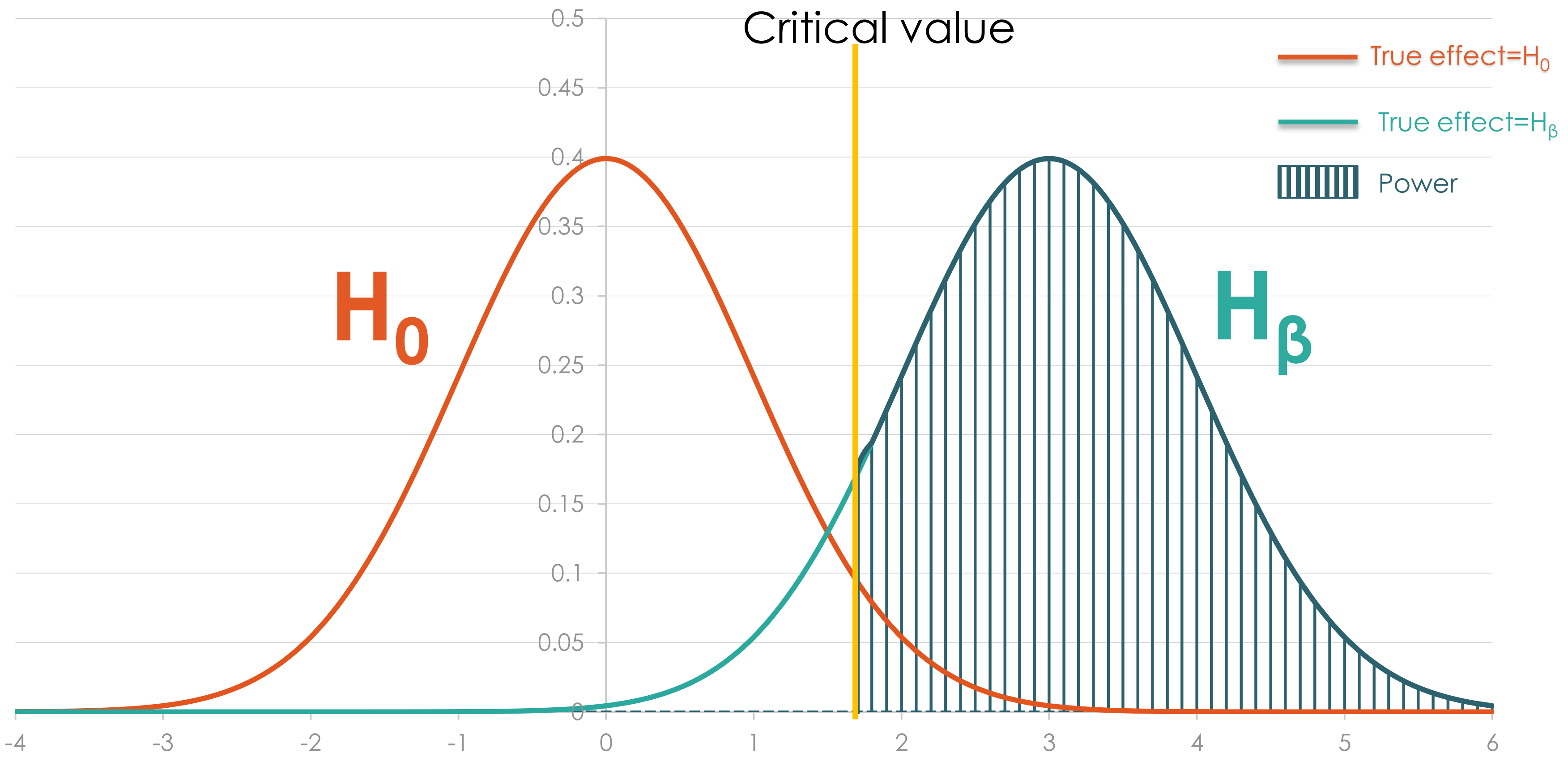
Outline

- Introduction
- Hypothesis testing
- **What influences power?**
- Power in clustered designs
- Calculating power in practice

What influences power?

Individual level randomization

What influences power?

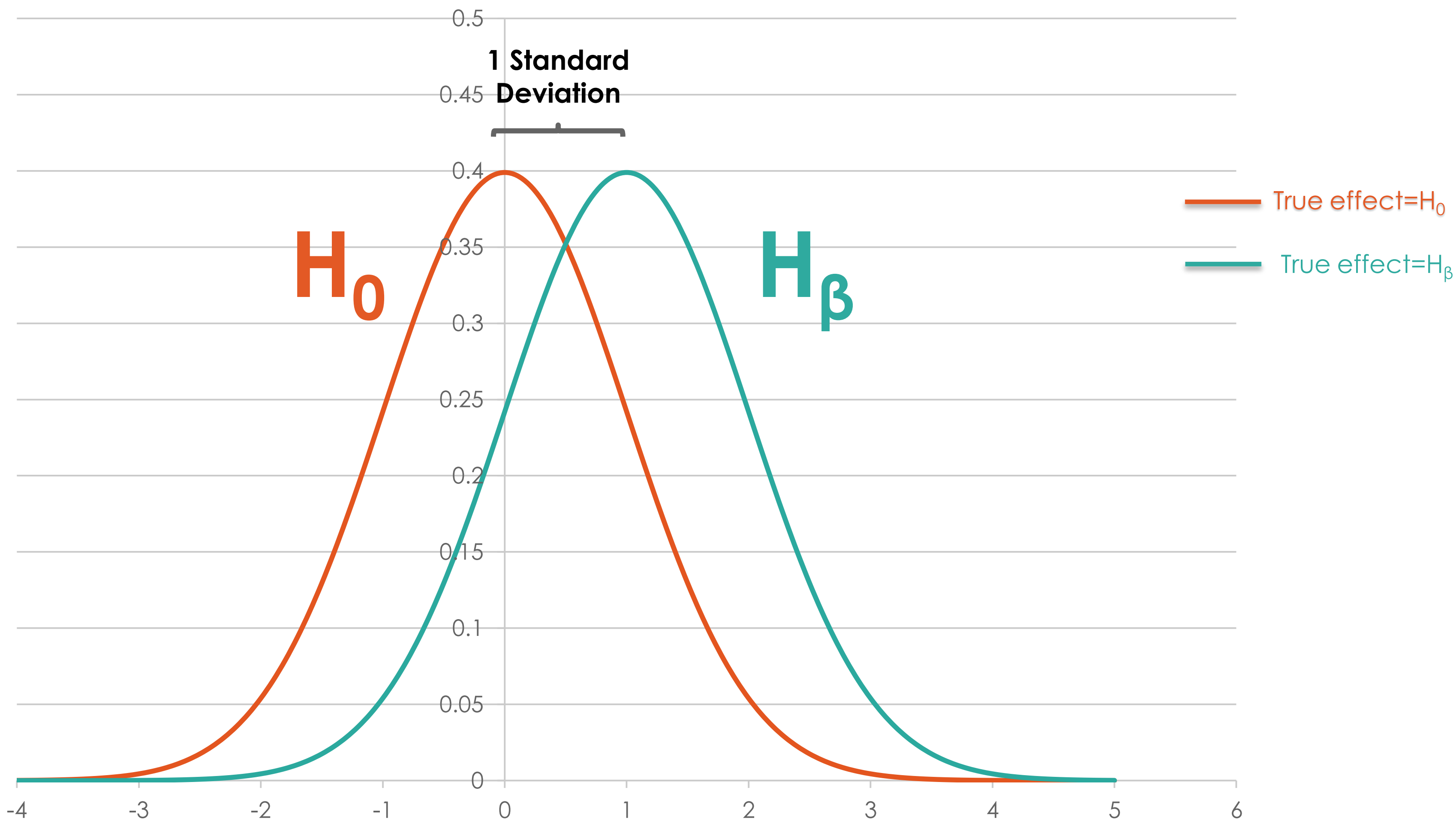


Move overlap of curves, less power, what influences overlap of these curves?

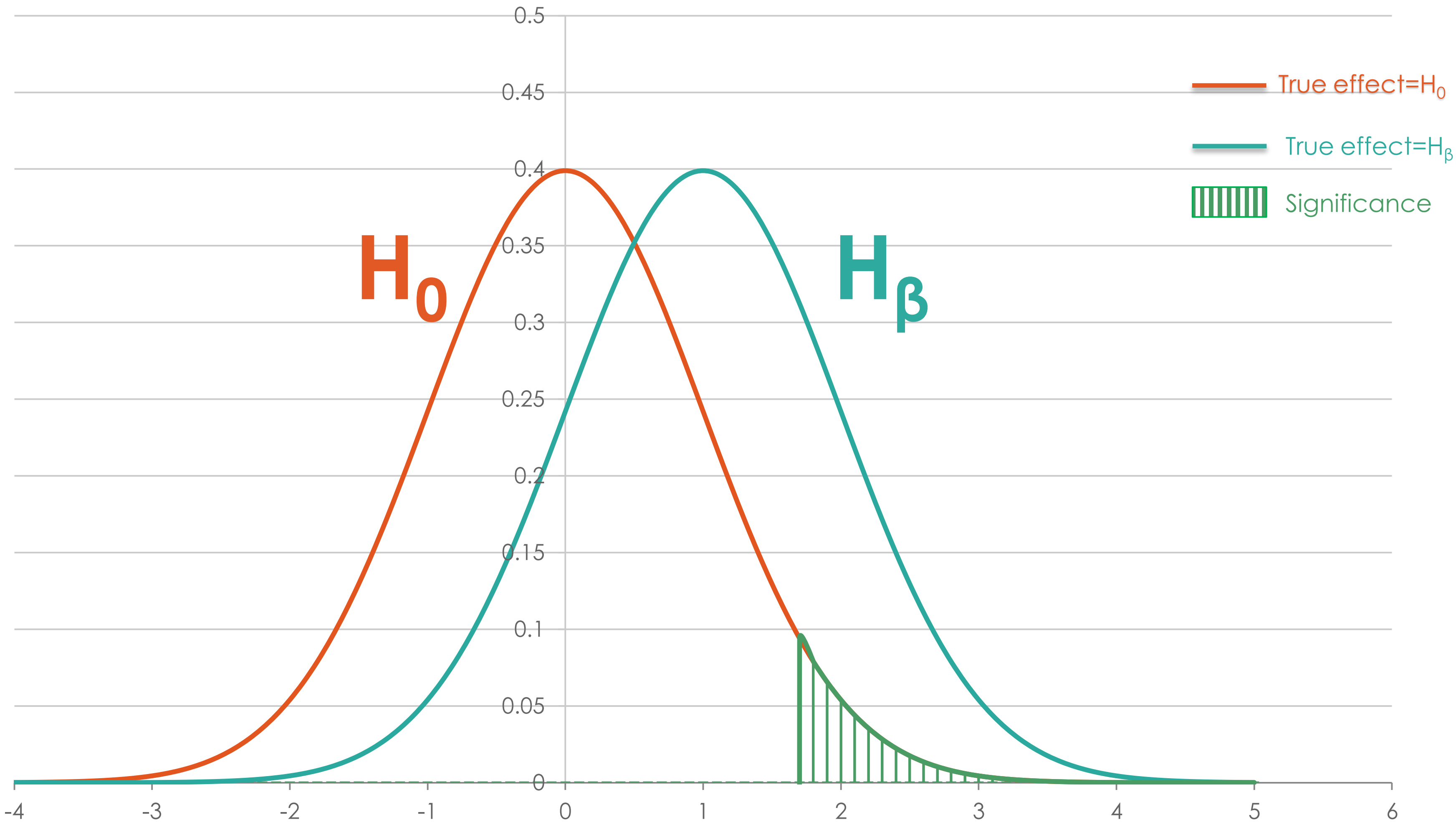
Power: main ingredients

1. Effect Size

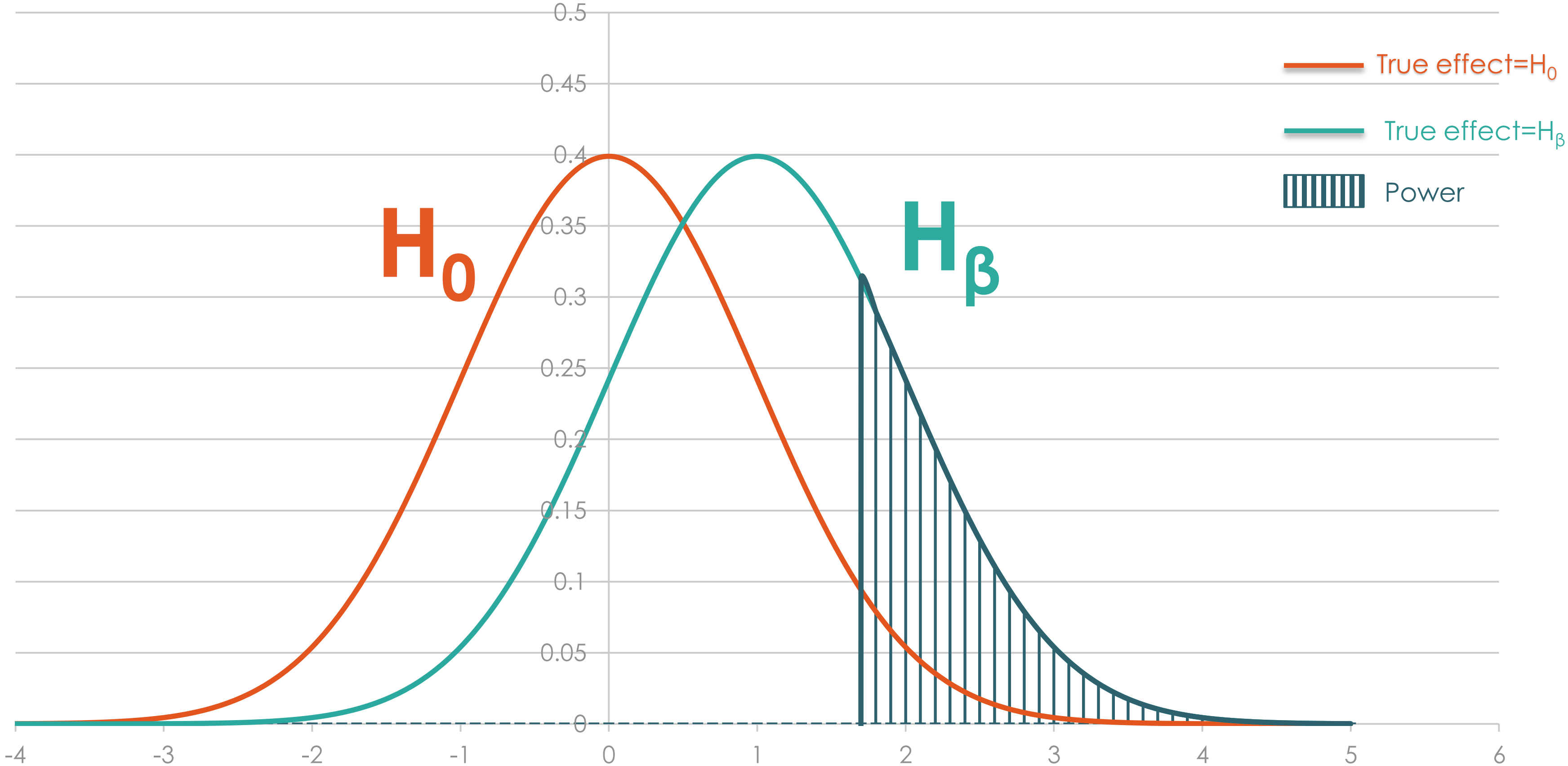
Effect Size: 1*SD



Effect Size = 1*SD

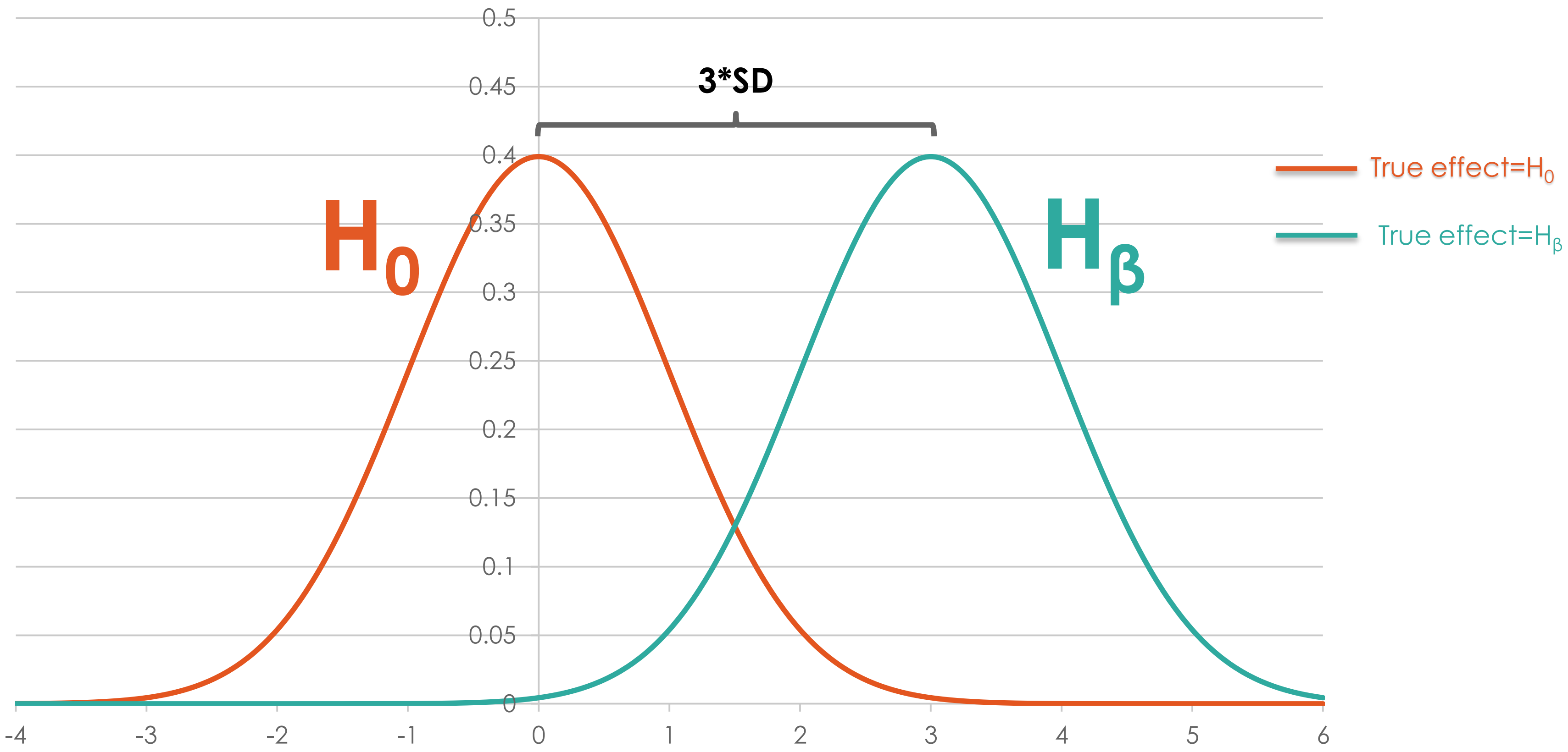


Power: 26%
If the true impact was 1*SD…



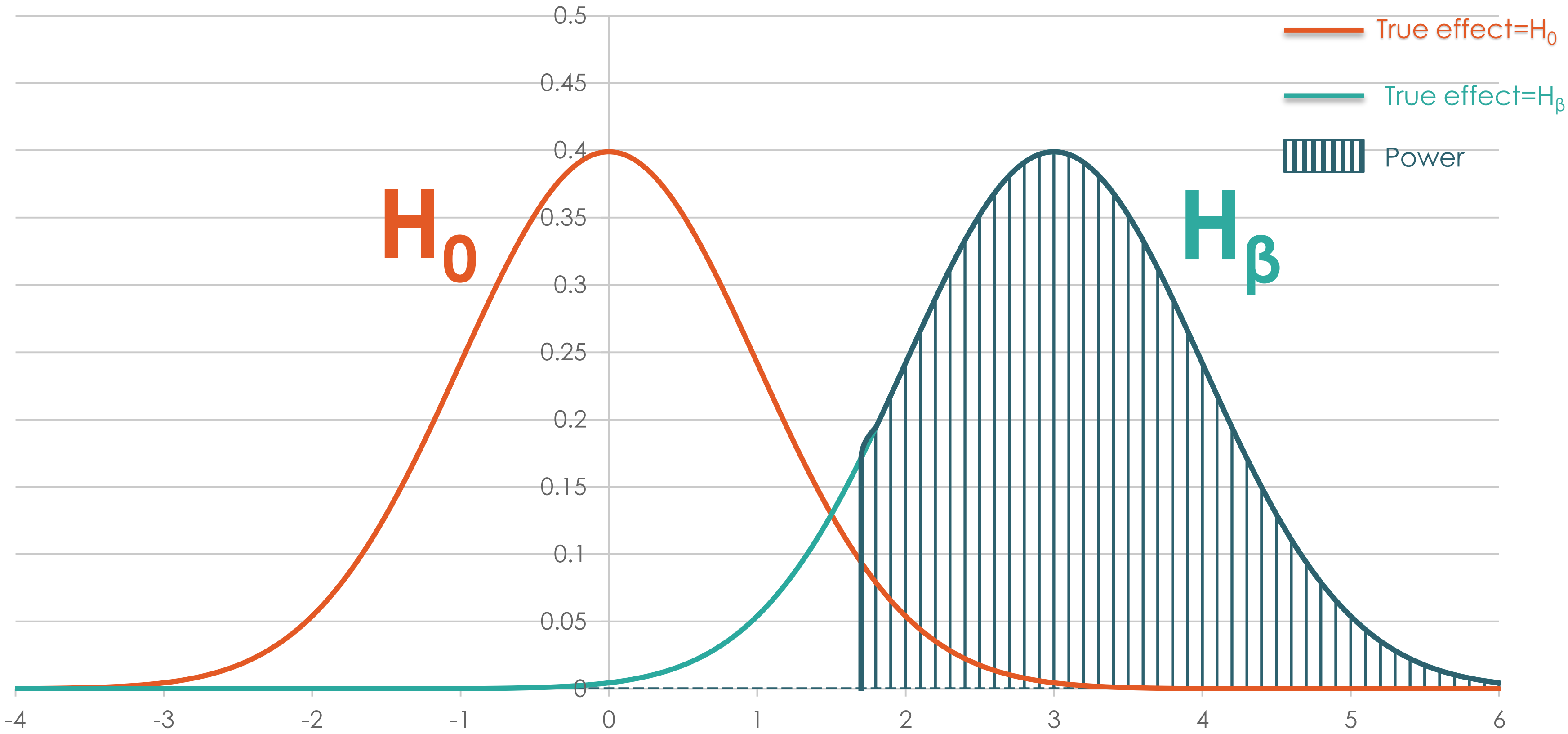
The Null Hypothesis would be rejected only 26% of the time

Effect Size: 3*SD



Bigger hypothesized effect size → distributions farther apart

Effect size 3*SD: Power= 91%

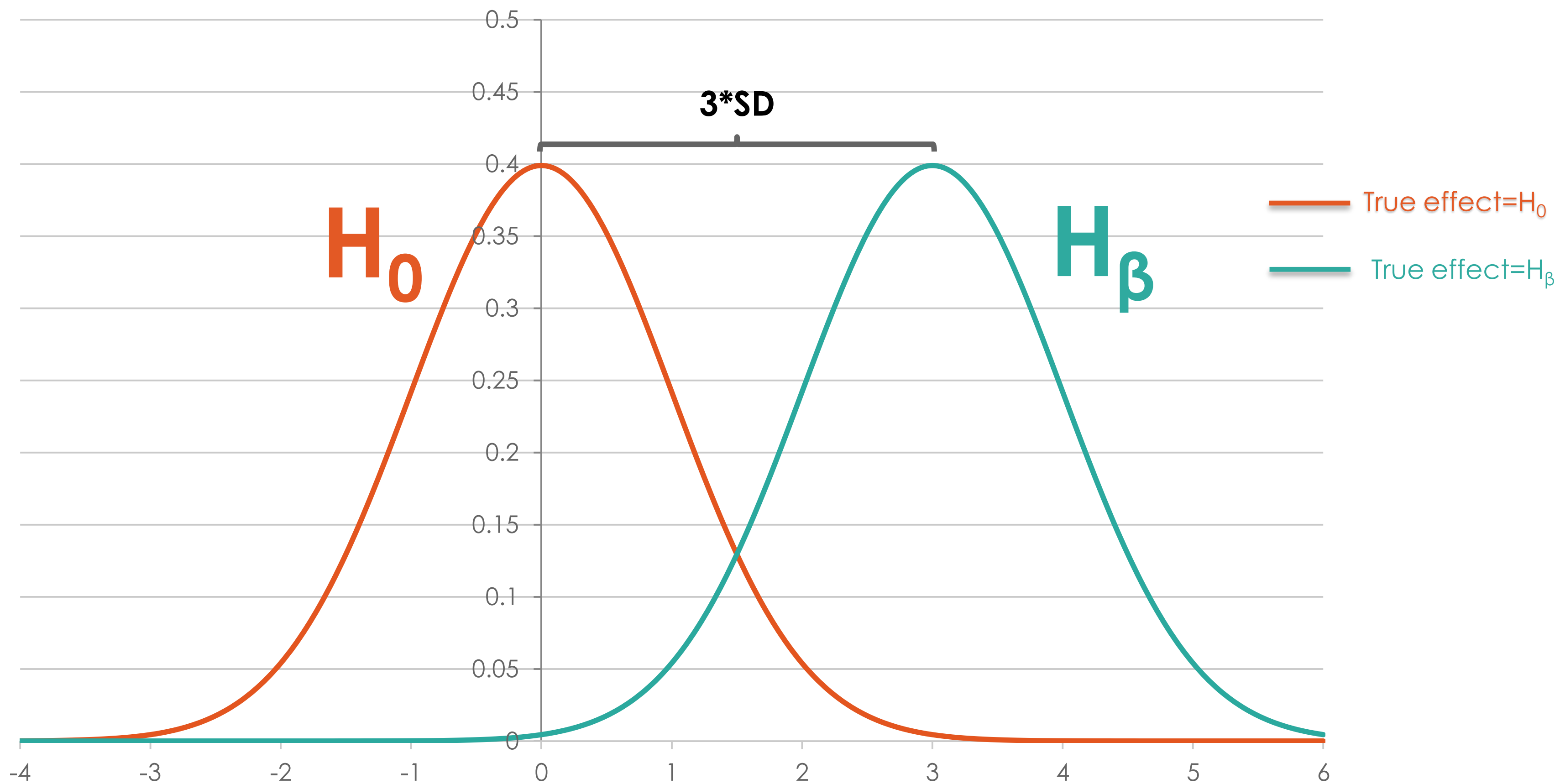


Bigger Effect size means more power

Effect size and take-up

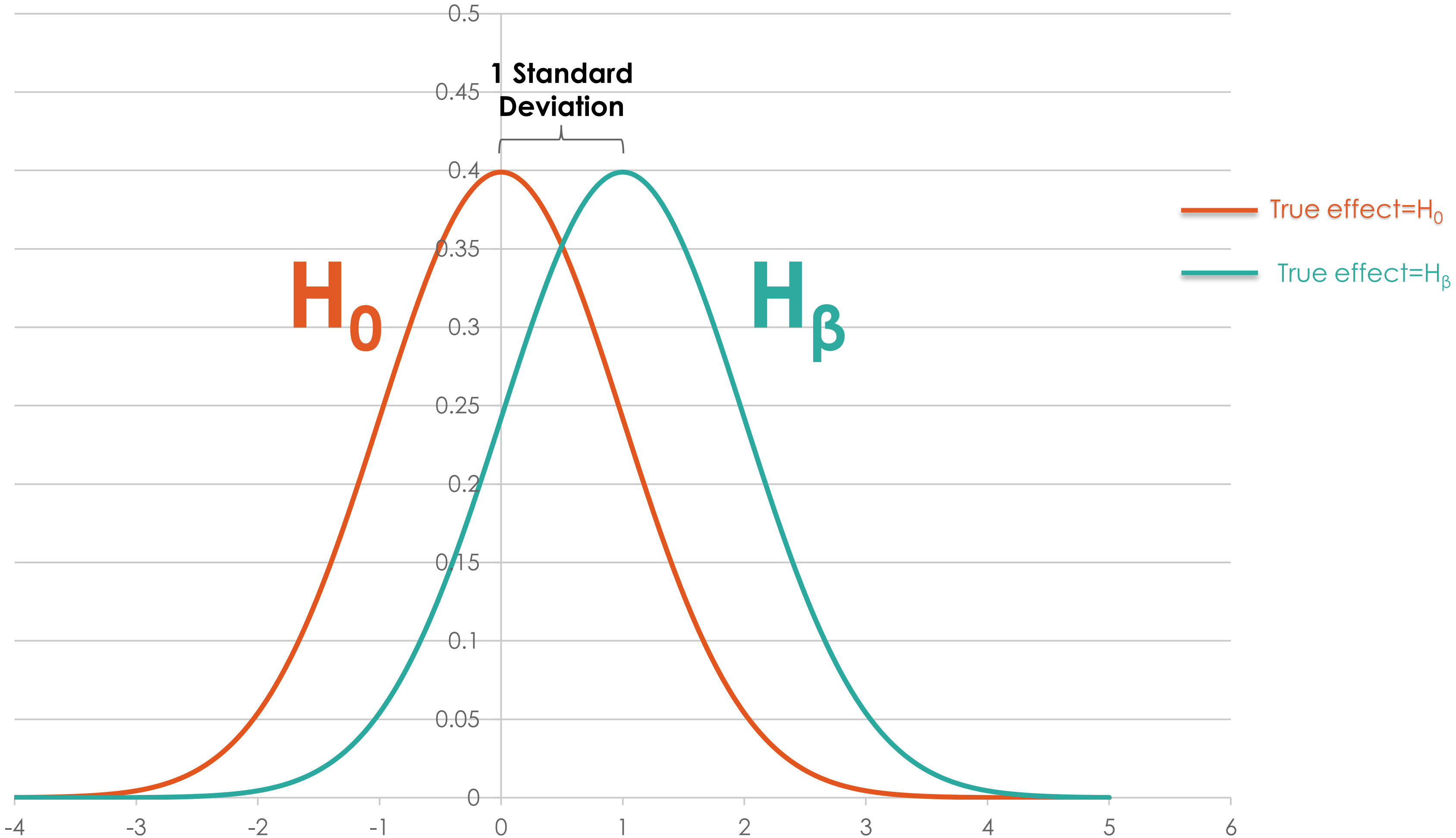
- Let’s say we believe the impact on our participants is “3 SD”
- What happens if take up is 1/3?
- Let’s show this graphically

Effect Size: 3*SD

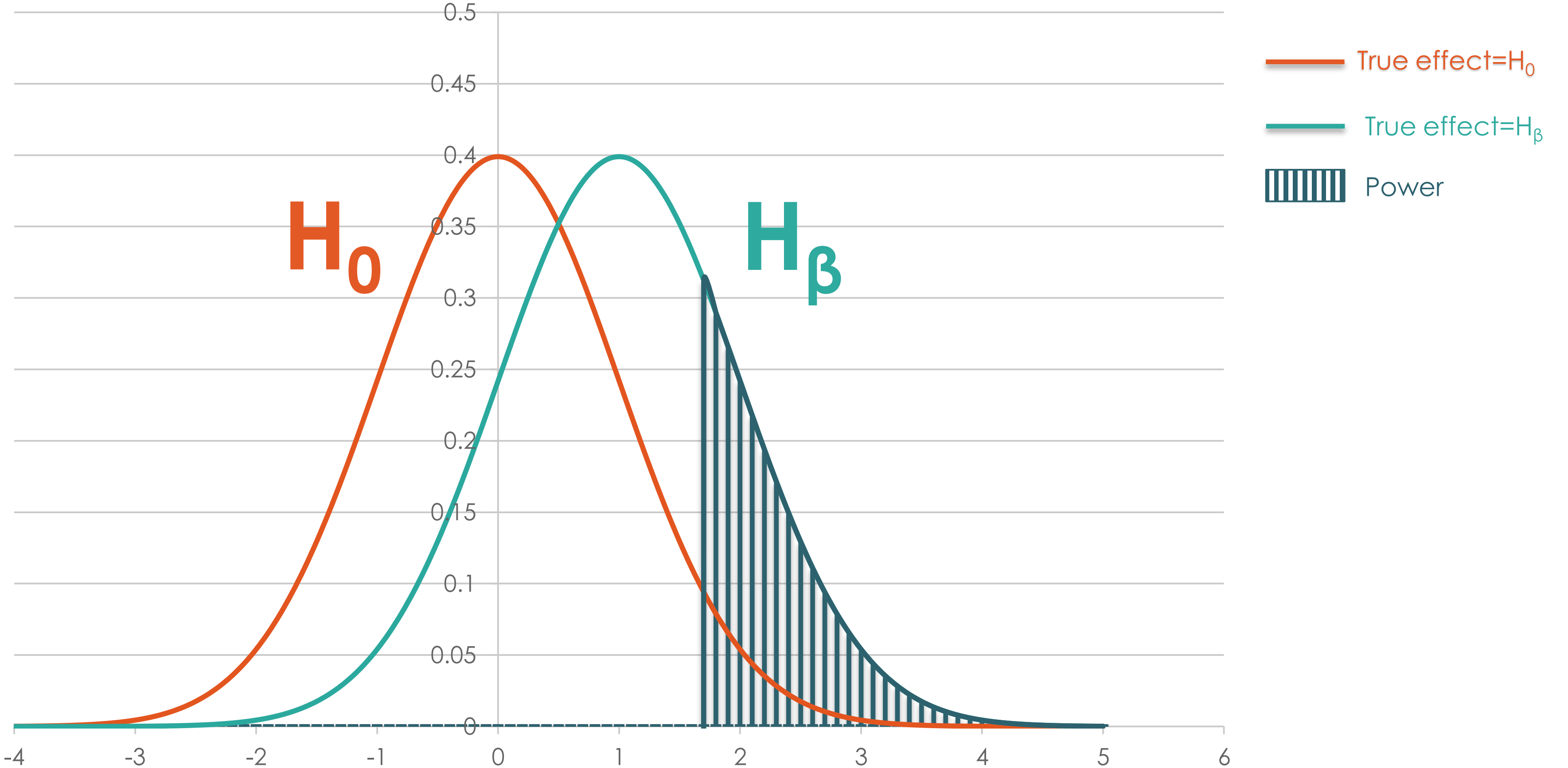


Let's say we believe the impact on our participants is "3SD"

Take up is 33%. Effect size is 1/3rd



Back to: Power = 26%

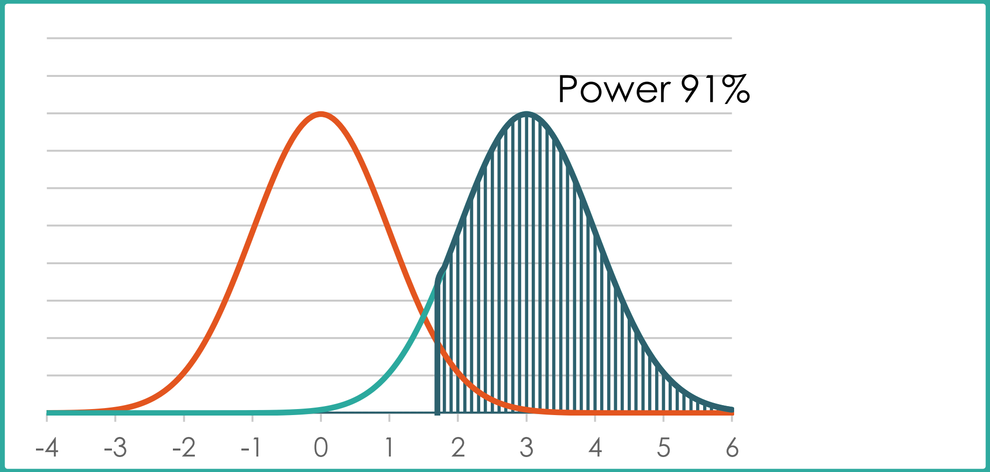


Take-up is reflected in the effect size

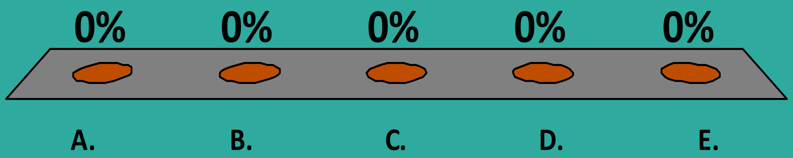
Power: main ingredients

1. Effect Size
2. Sample Size

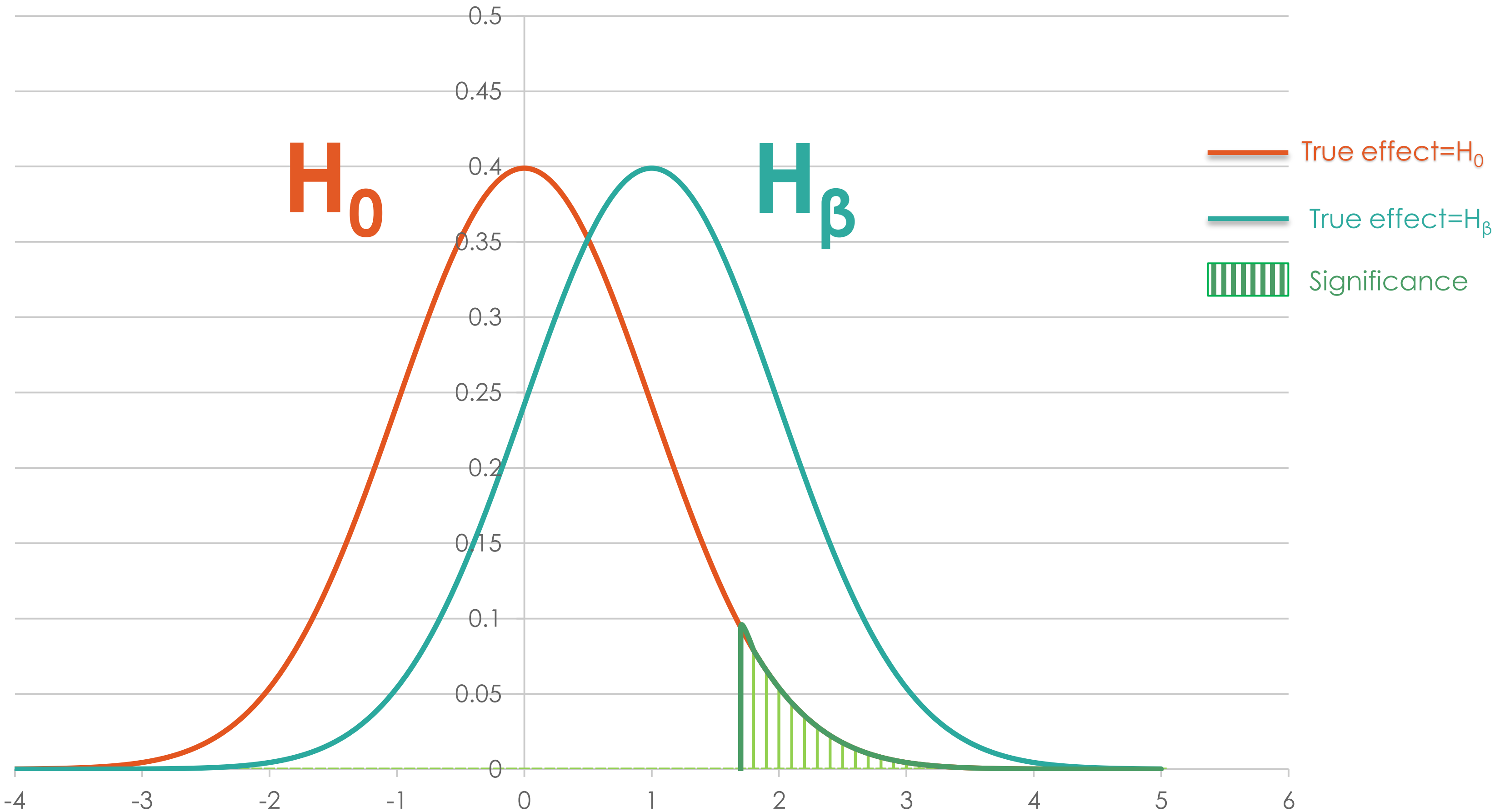
Increasing sample size will …



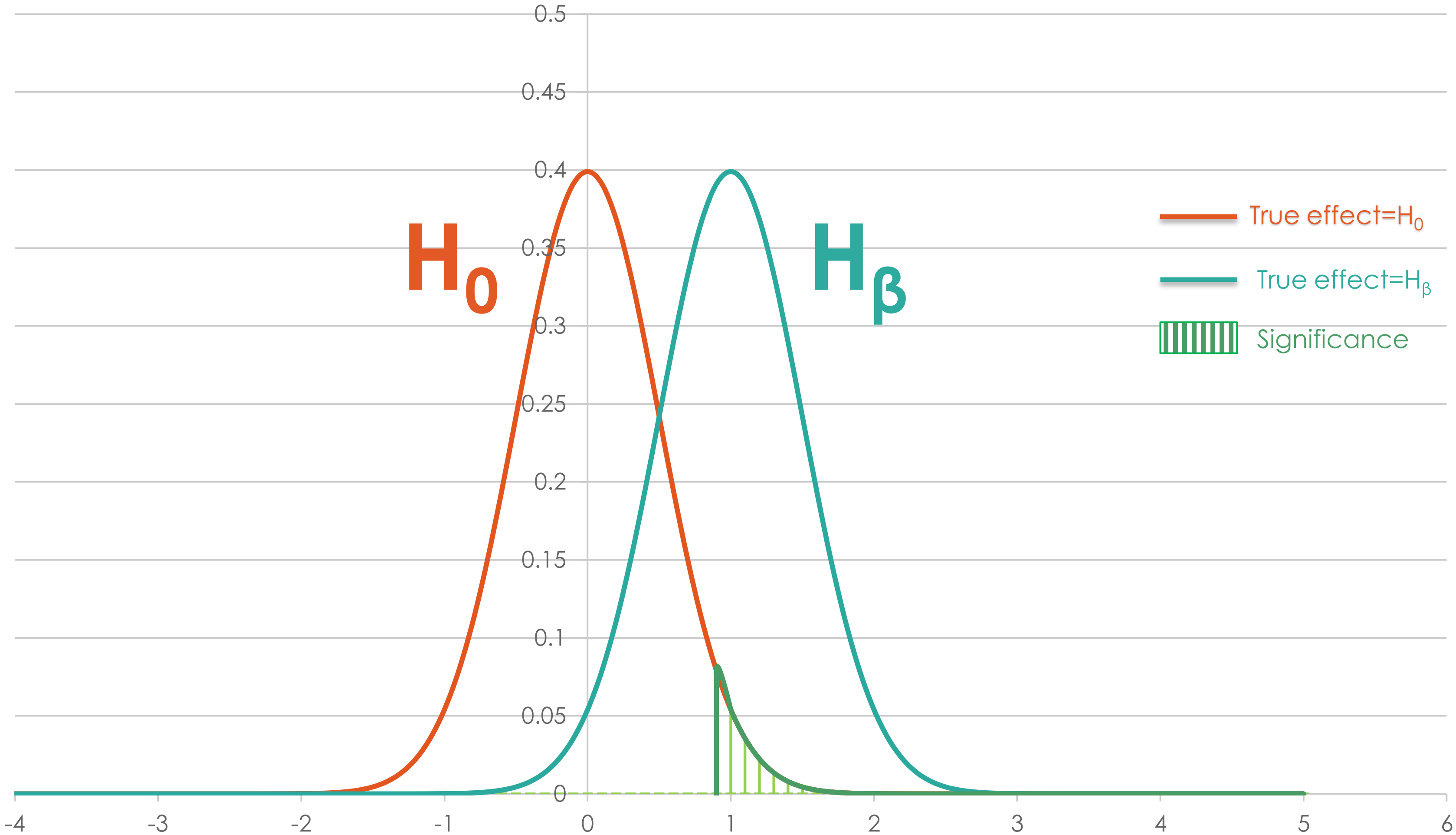
A. Move curves further apart
B. Move curves closer together
C. Make curves fatter
D. Make curves narrower
E. Don't know



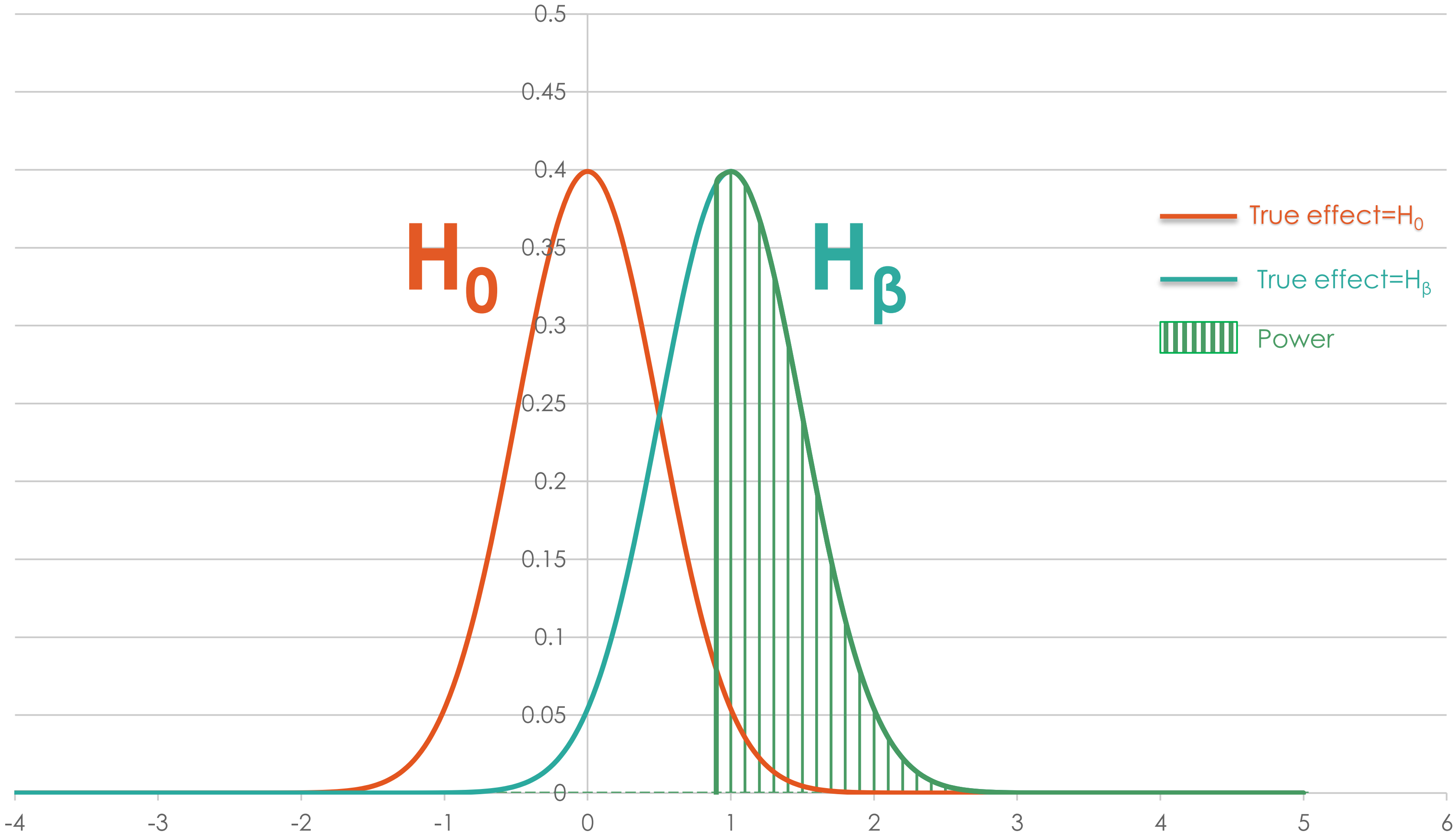
Power: Effect size = 1 SD
Sample size = 1,000



Power: Effect size= 1SD
Sample size = 4,000

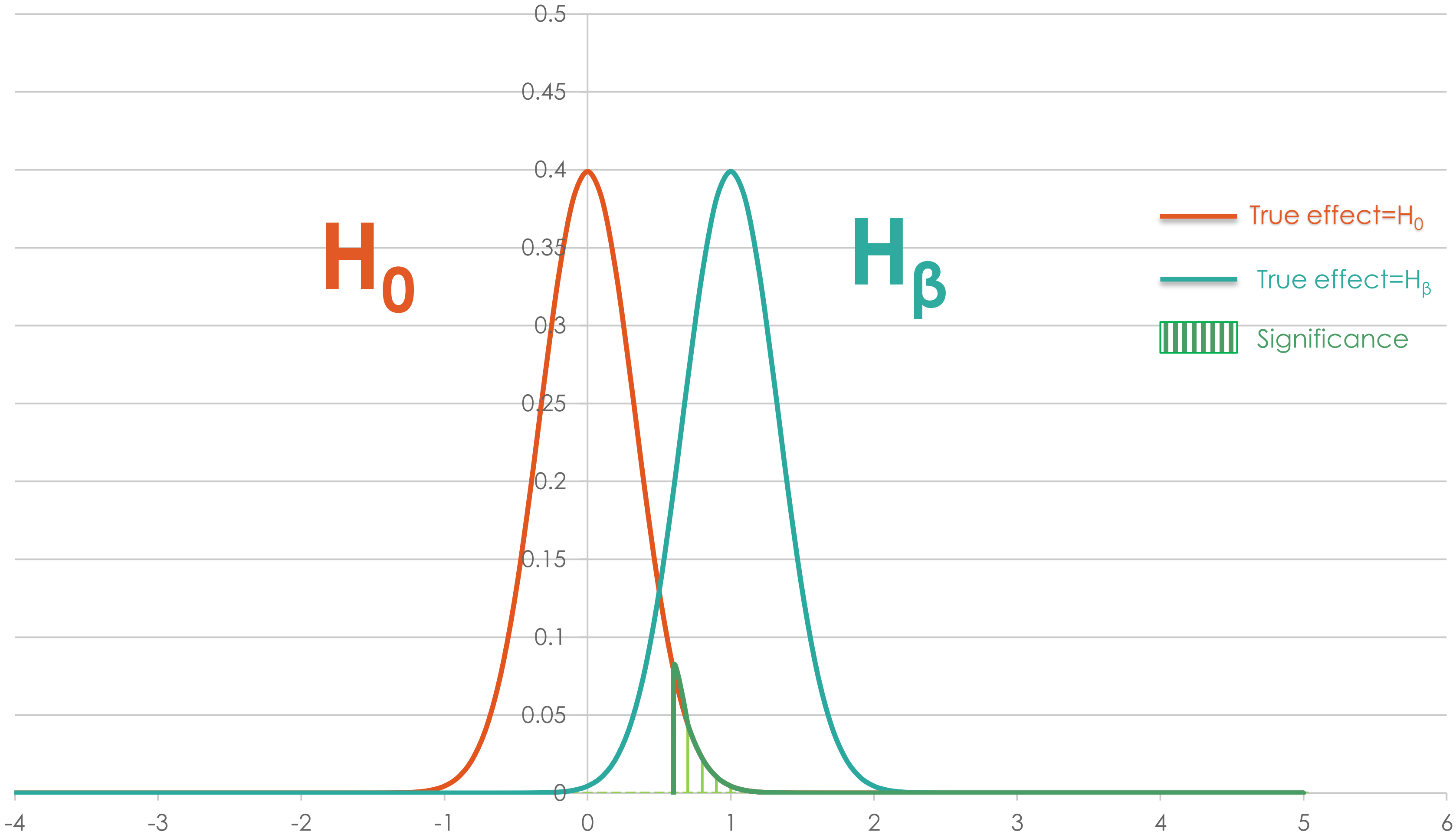


Power: 64%

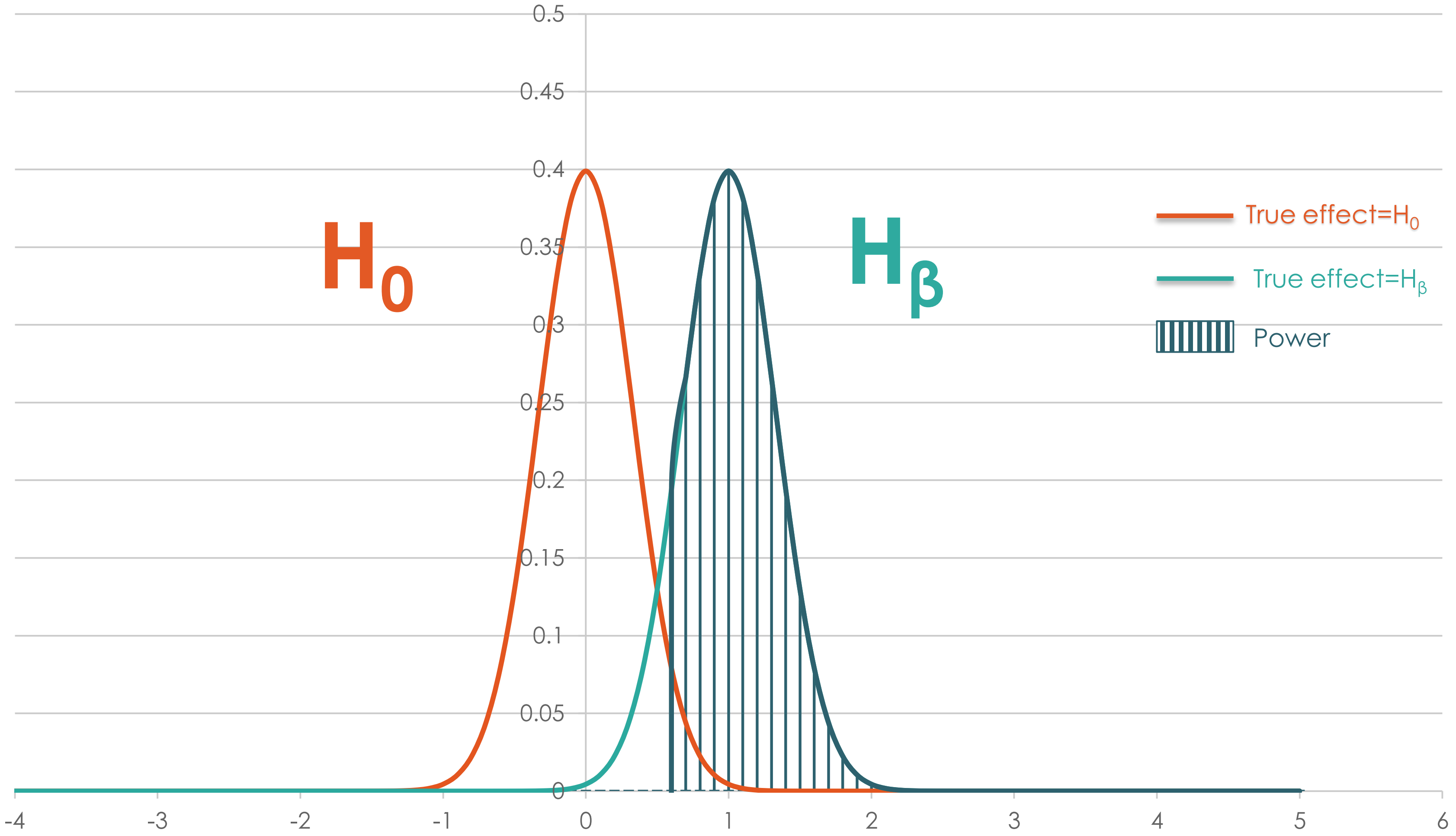


H0
Hβ
True effect=H0
True effect=Hβ
Power
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-4
-3
-2
-1
0
1
2
3
4
5
6

Power: Sample size = 9,000



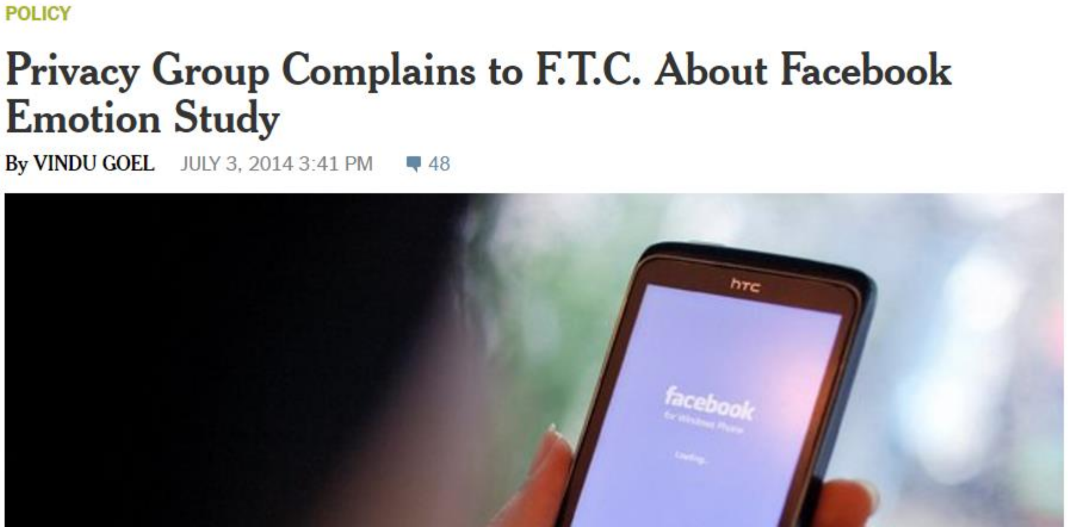
Power: 91%



0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-4
-3
-2
-1
0
1
2
3
4
5
6
H_0
H_β
True effect=H_0
True effect=H_β
Power

Sample Size = 700,000?

- Controversial Facebook study on emotional contagion:
- Less positive content on News Feed
  → Fewer positive words in status updates
- How big was the effect? Tiny, *but significant:*
  – % Positive posts falls from ~5.2% to ~5.1%


POLICY

Privacy Group Complains to F.T.C. About Facebook Emotion Study

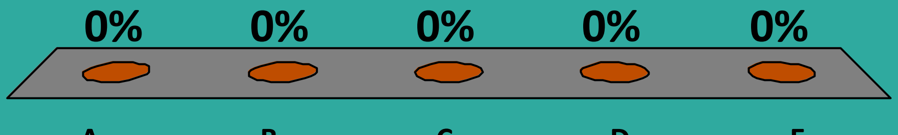
By VINDU GOEL JULY 3, 2014 3:41 PM 48

Power: main ingredients

1. Effect Size
2. Sample Size
3. Variance

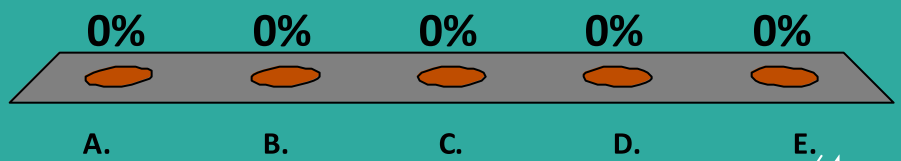
What will increased variation in the underlying population do to our estimates?

A. Increase risk of bias
B. Reduce risk of bias
C. Increase precision of estimate
D. Reduce precision of estimate
E. Will not change estimates

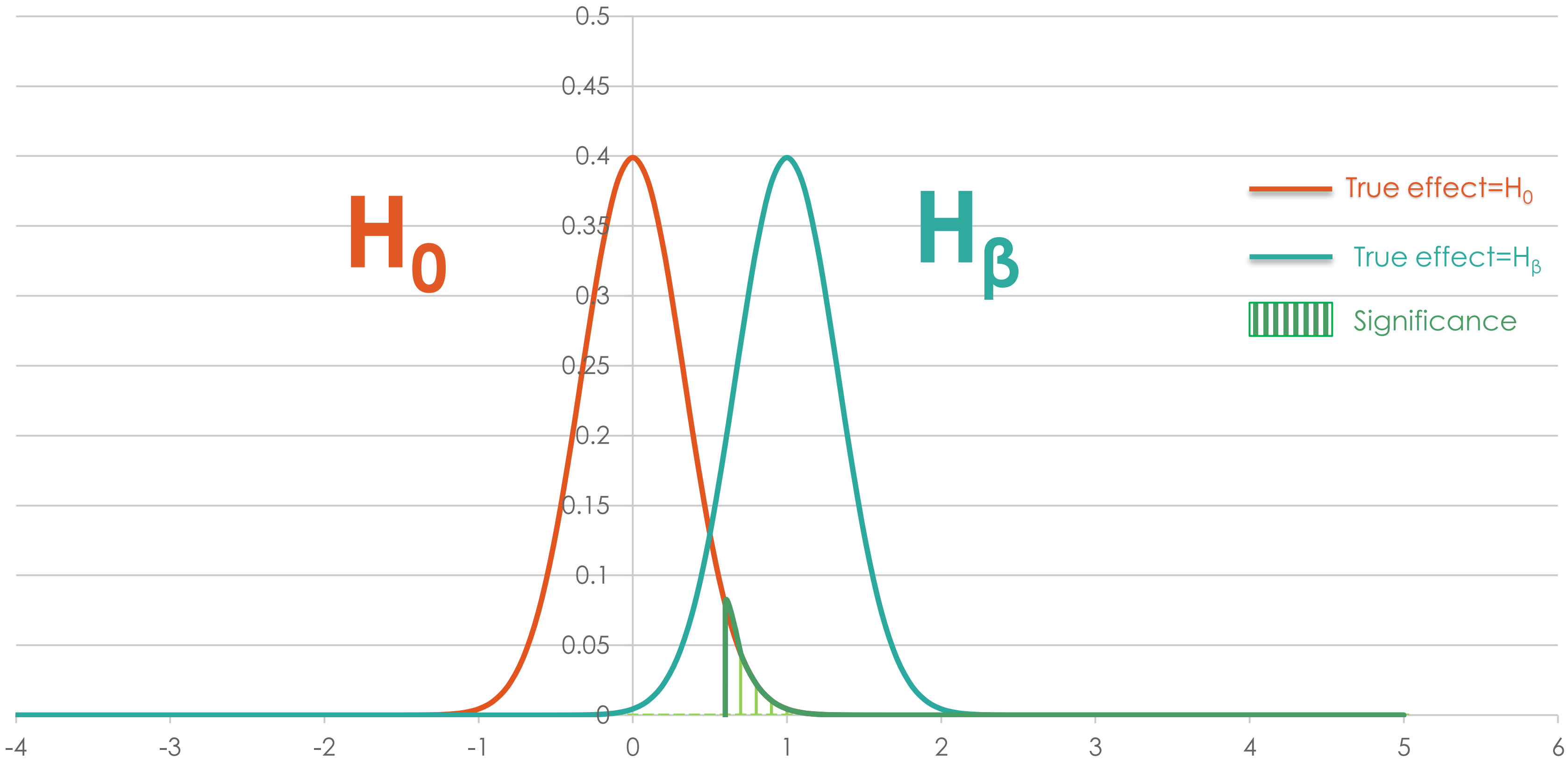


What does increased variation in population do to our distribution of estimates curves?

A. Move them further apart
B. Move them closer
C. Make them fatter
D. Make them thinner
E. Don’t know

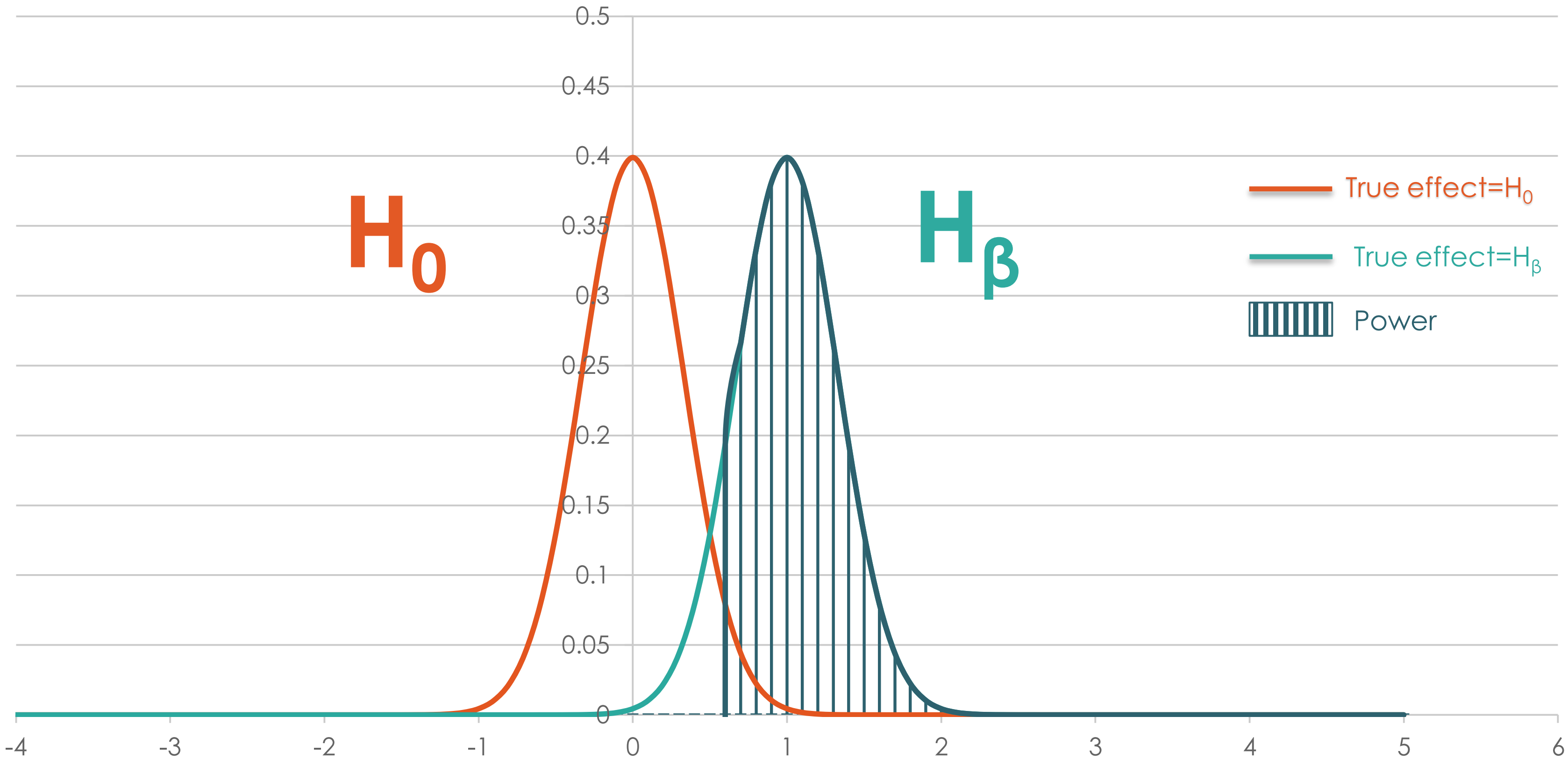


Low variance sample



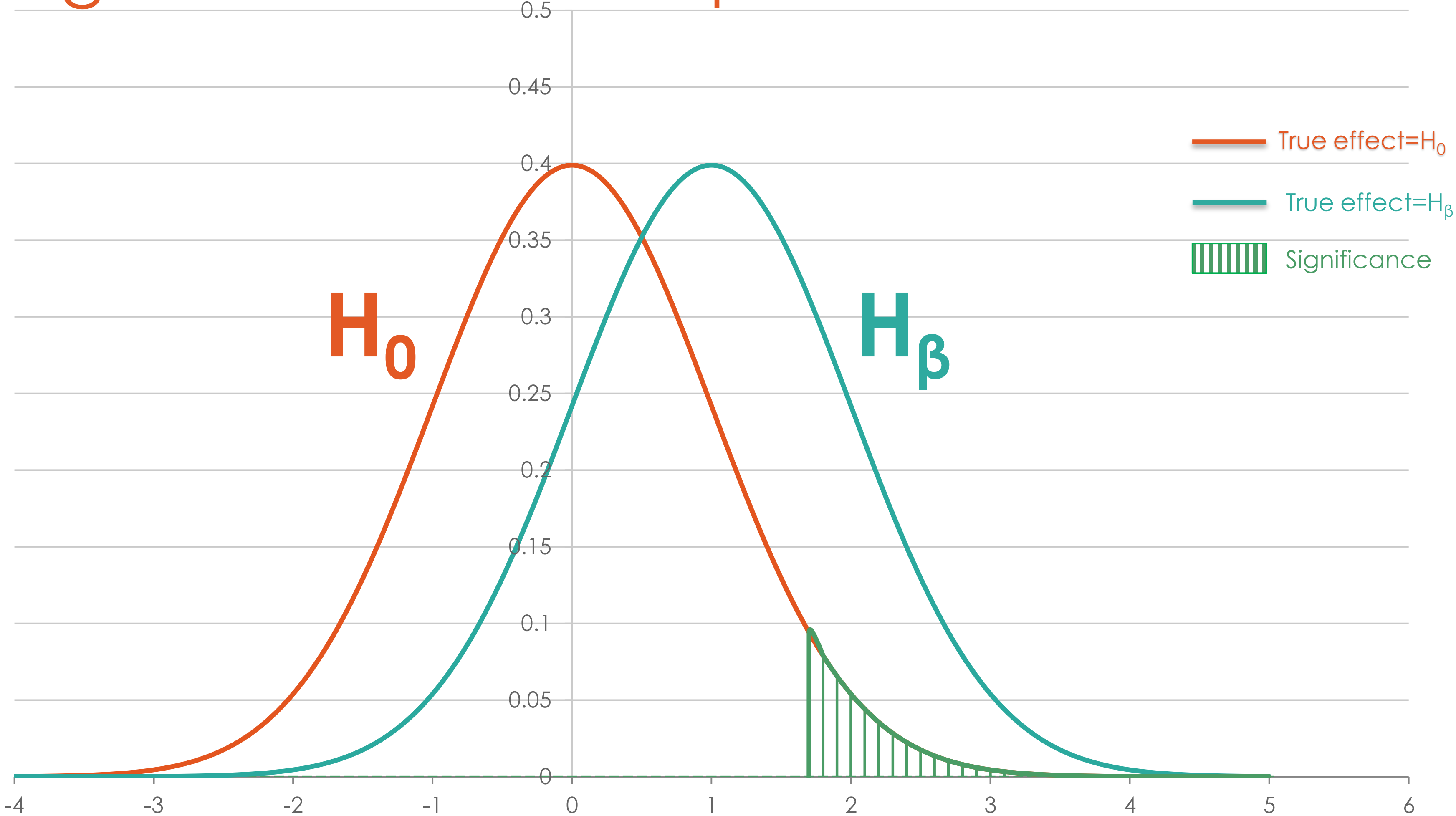
Estimates will be more tightly clustered

Low variance sample



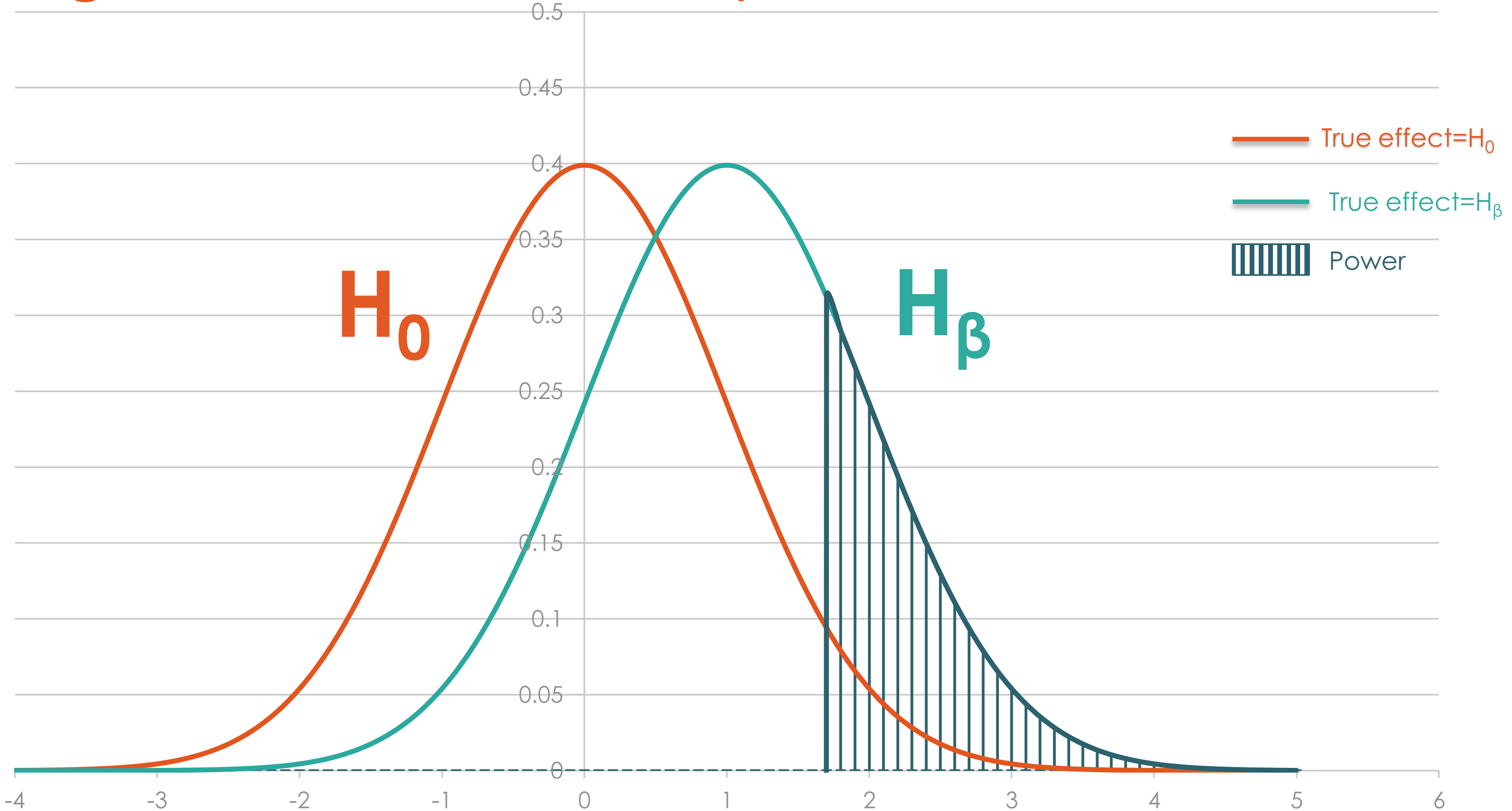
Estimates will be more tightly clustered ➔ Higher power

Higher variance sample



Estimates will be more dispersed

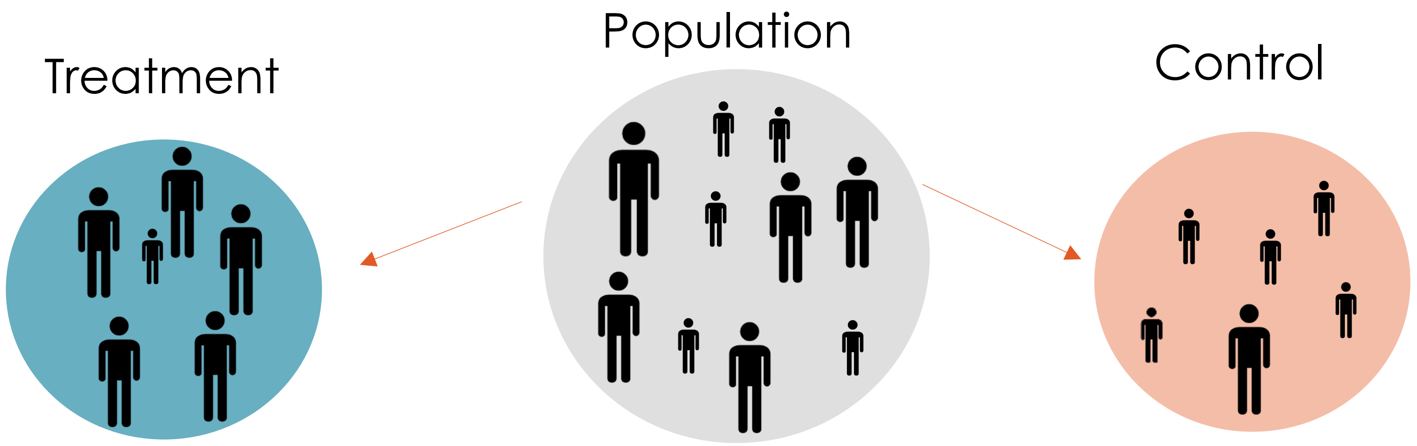
Higher variance sample



Estimates will be more dispersed ➔ Lower power

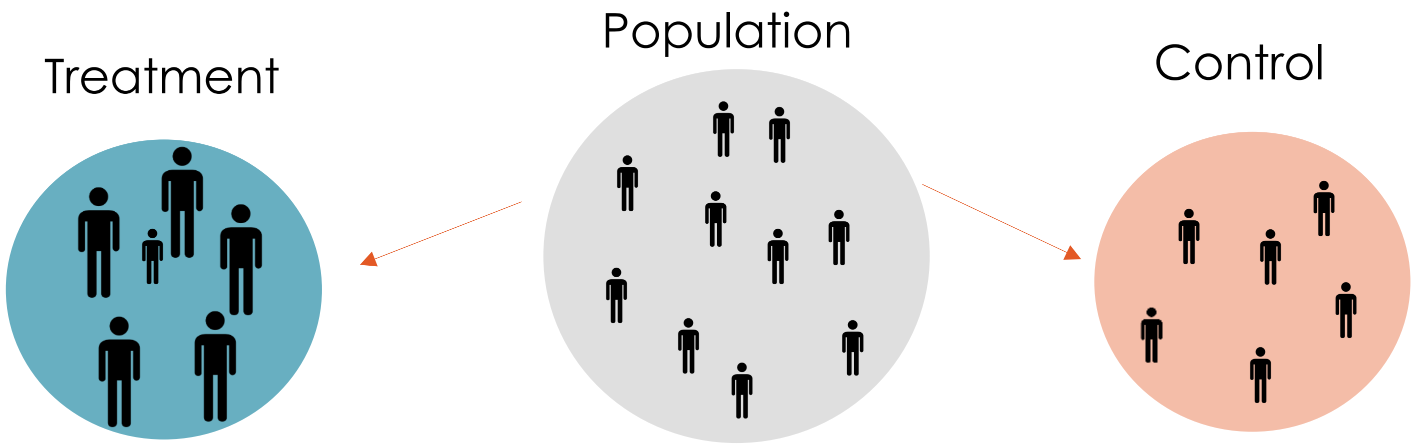
Variance and power: intuition

- Our program seeks to increase child height
- There is a lot of variation in height in the population
- At endline children in treatment are taller than in control
- Is this because we happened to start with taller children? or because the program worked?
- We need a big sample to sort this out



Variance and power: intuition

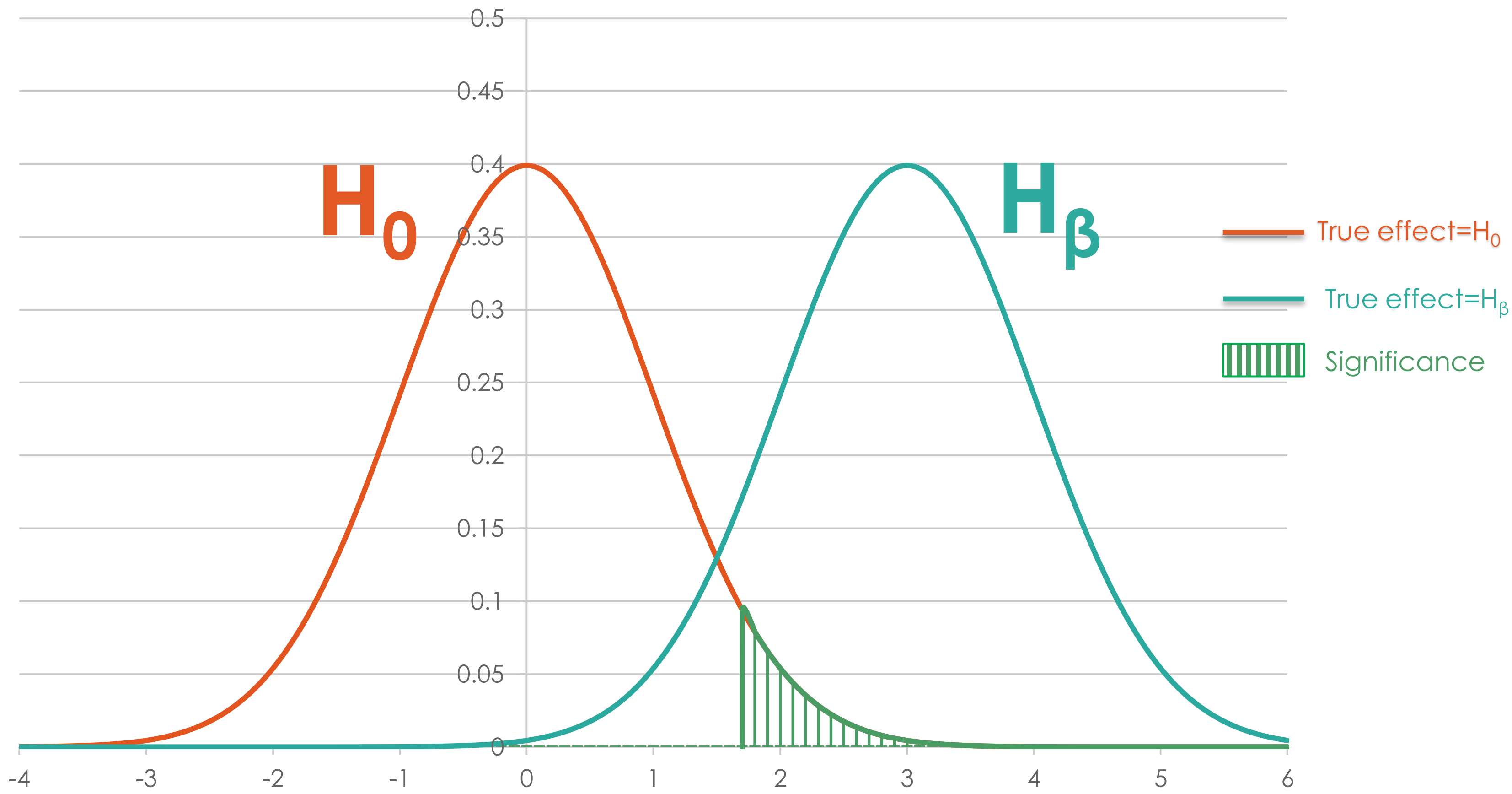
- If everyone in the underlying population was of similar height at the start, it would be easy to sort this out
  - we would need a smaller sample (or have more power with a given sample)
  - the variation we see at the end between treatment and control must be due to the program



Power: main ingredients

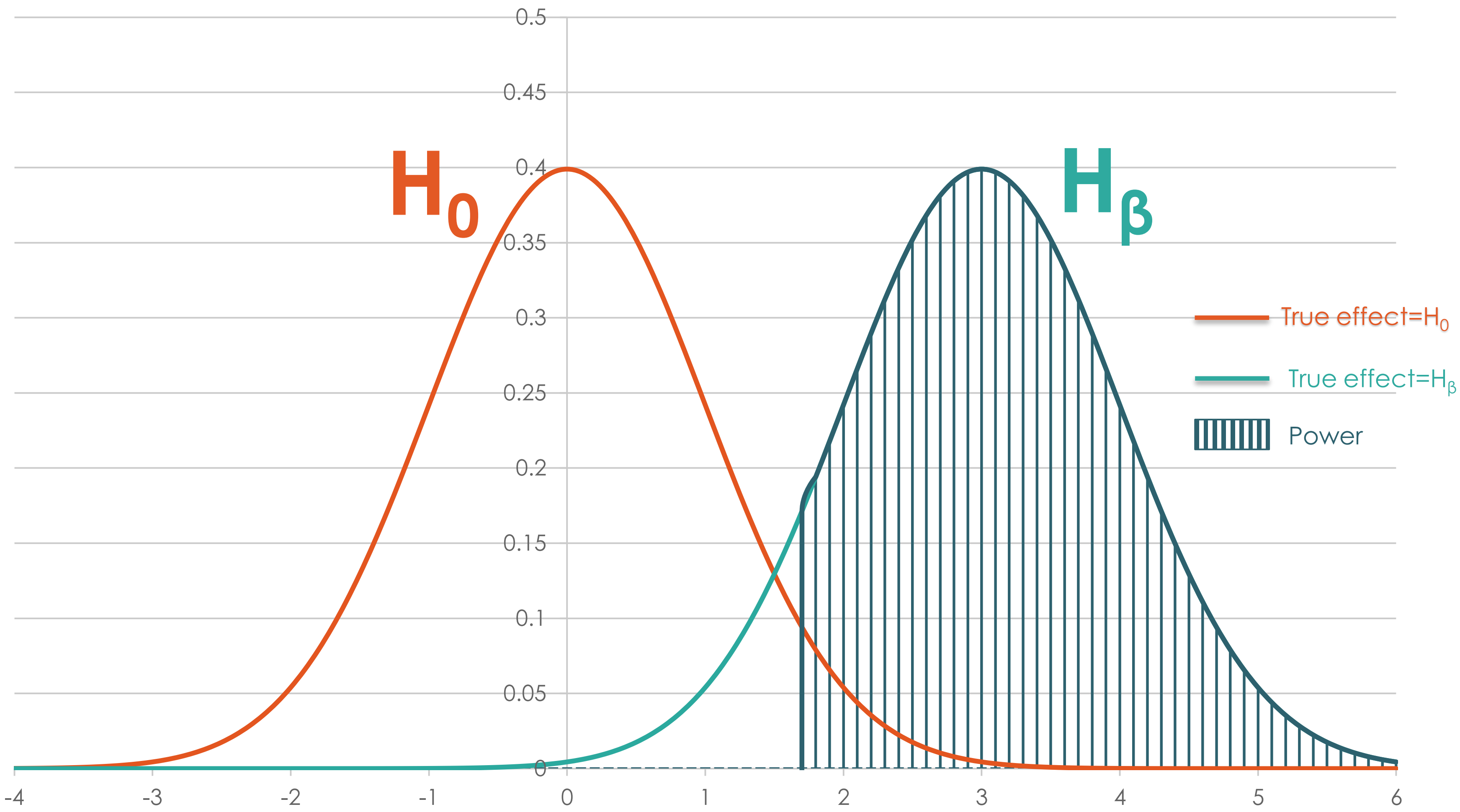
1. Effect Size
2. Sample Size
3. Variance
4. Proportion of sample in T vs. C

Sample split: 50% C, 50% T



Equal split gives distributions that are the same "fatness"

Power: 91%

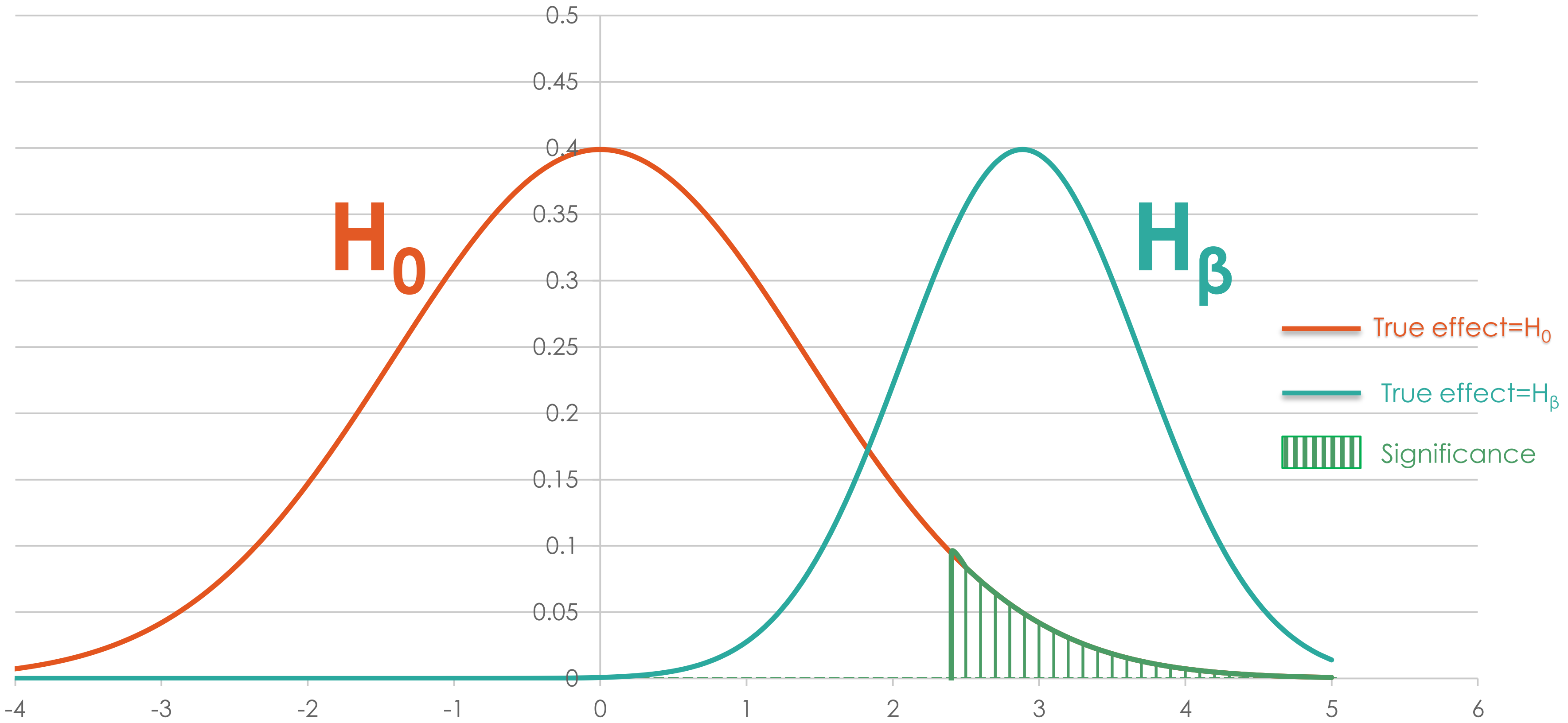


0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
H_0
H_β
True effect=H_0
True effect=H_β
Power
-4
-3
-2
-1
0
1
2
3
4
5
6

If it’s not 50-50 split?

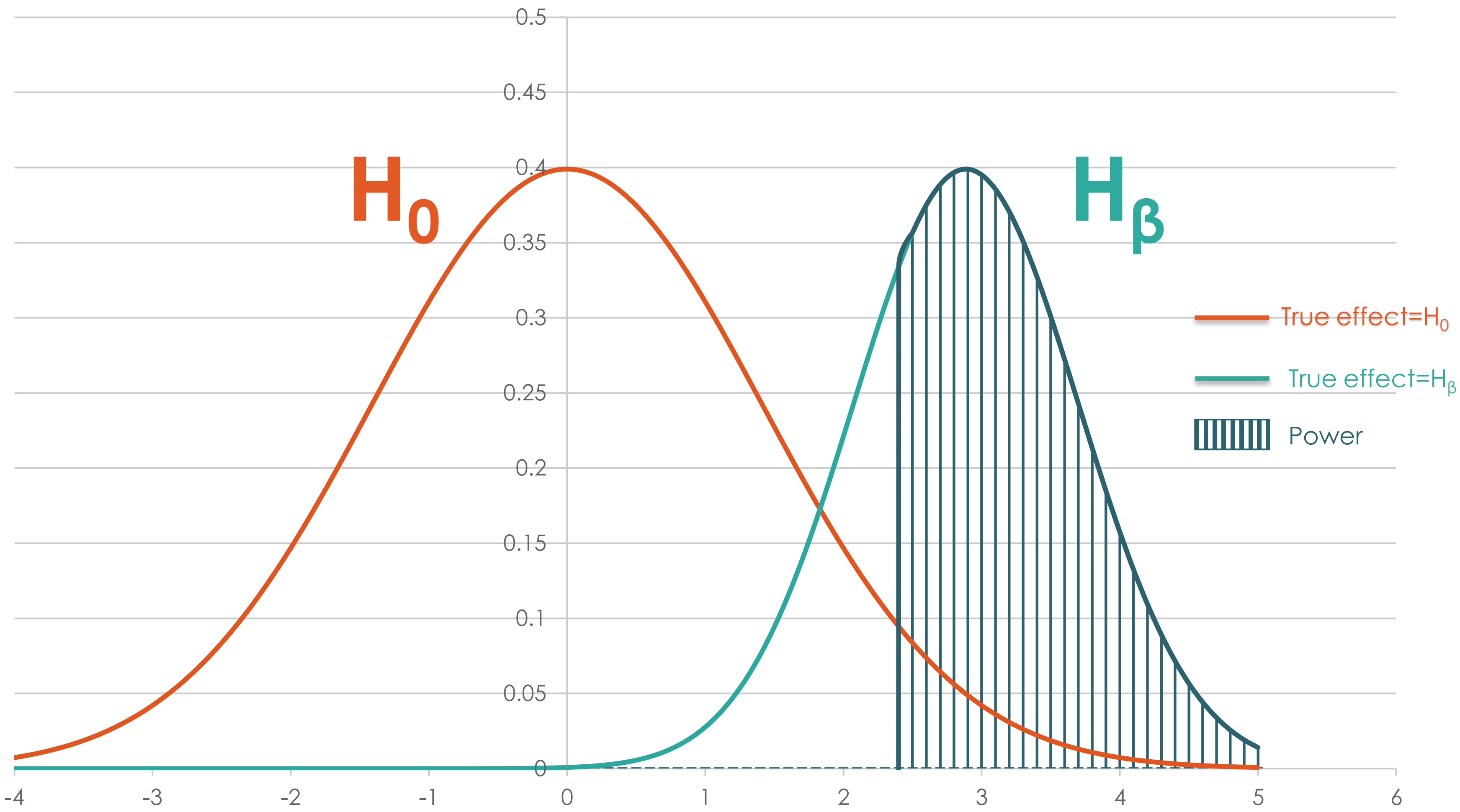
- What happens to the relative fatness if the split is not 50-50.
- Say 25-75?

Sample split: 25% C, 75% T



Uneven distributions, not efficient, i.e. less power

Power: 83%



H0
Hβ
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-4
-3
-2
-1
0
1
2
3
4
5
6
True effect=H0
True effect=Hβ
Power

Allocation ratio

- <u>Definition</u>: the fraction of the total sample allocated to the treatment group is the allocation ratio
- Usually, for a given sample size, power is maximized when half sample allocated to treatment, half to control
- Diminishing marginal benefit to precision from adding sample, so best to add equally

Allocation to T v C

$$sd(X_1 - X_2) = \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}$$

$$sd(X_1 - X_2) = \sqrt{\frac{1}{2} + \frac{1}{2}} = \sqrt{\frac{2}{2}} = 1$$

$$sd(X_1 - X_2) = \sqrt{\frac{1}{3} + \frac{1}{1}} = \sqrt{\frac{4}{3}} = 1.15$$

Power equation: MDE

Significance Level

Variance

Effect Size

Power

$$EffectSize = \left(t_{(1-\kappa)} + t_{\alpha}\right) * \sqrt{\frac{1}{P(1-P)}} * \sqrt{\frac{\sigma^2}{N}}$$

Proportion in Treatment

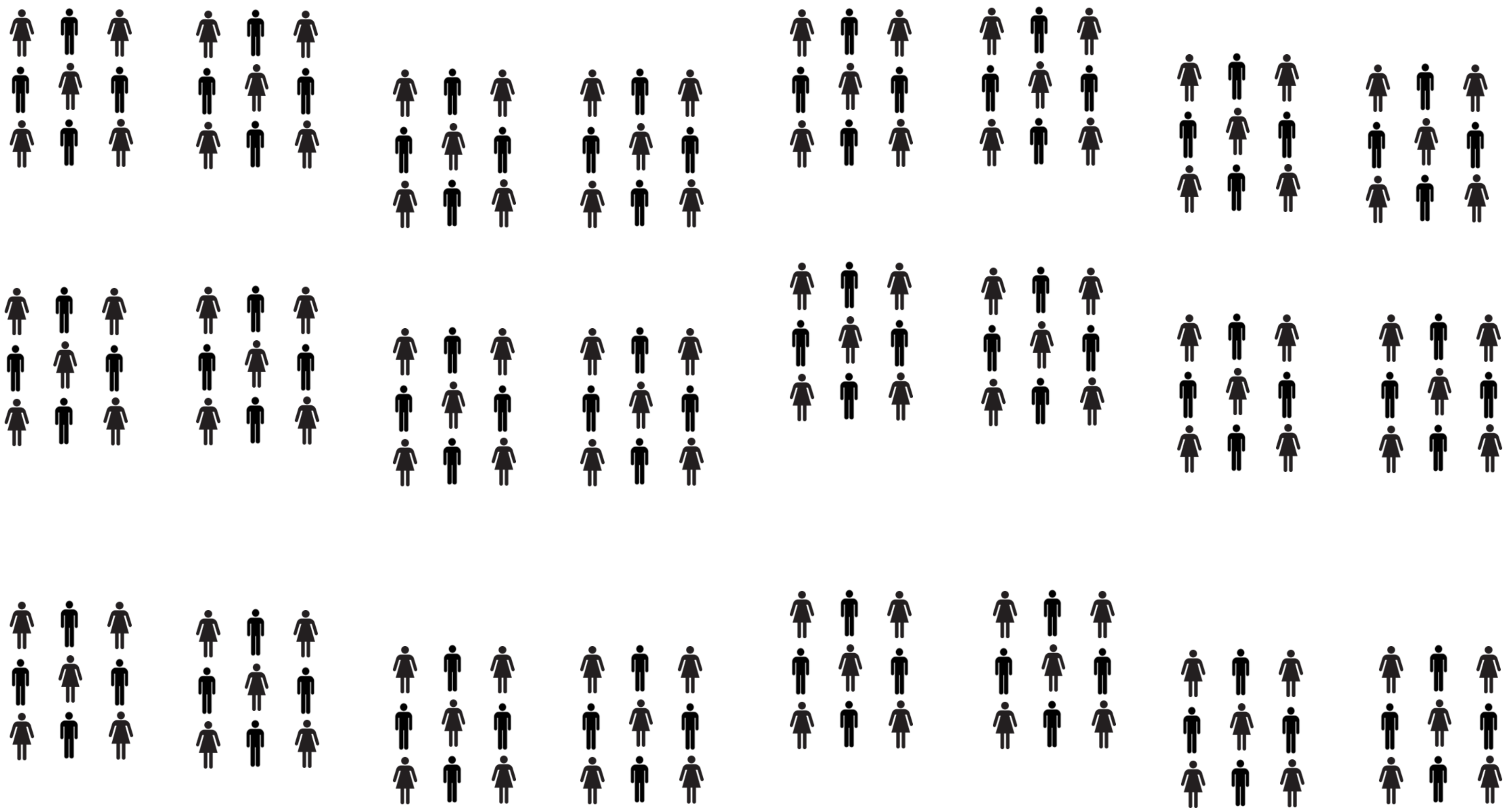
Sample Size

Outline

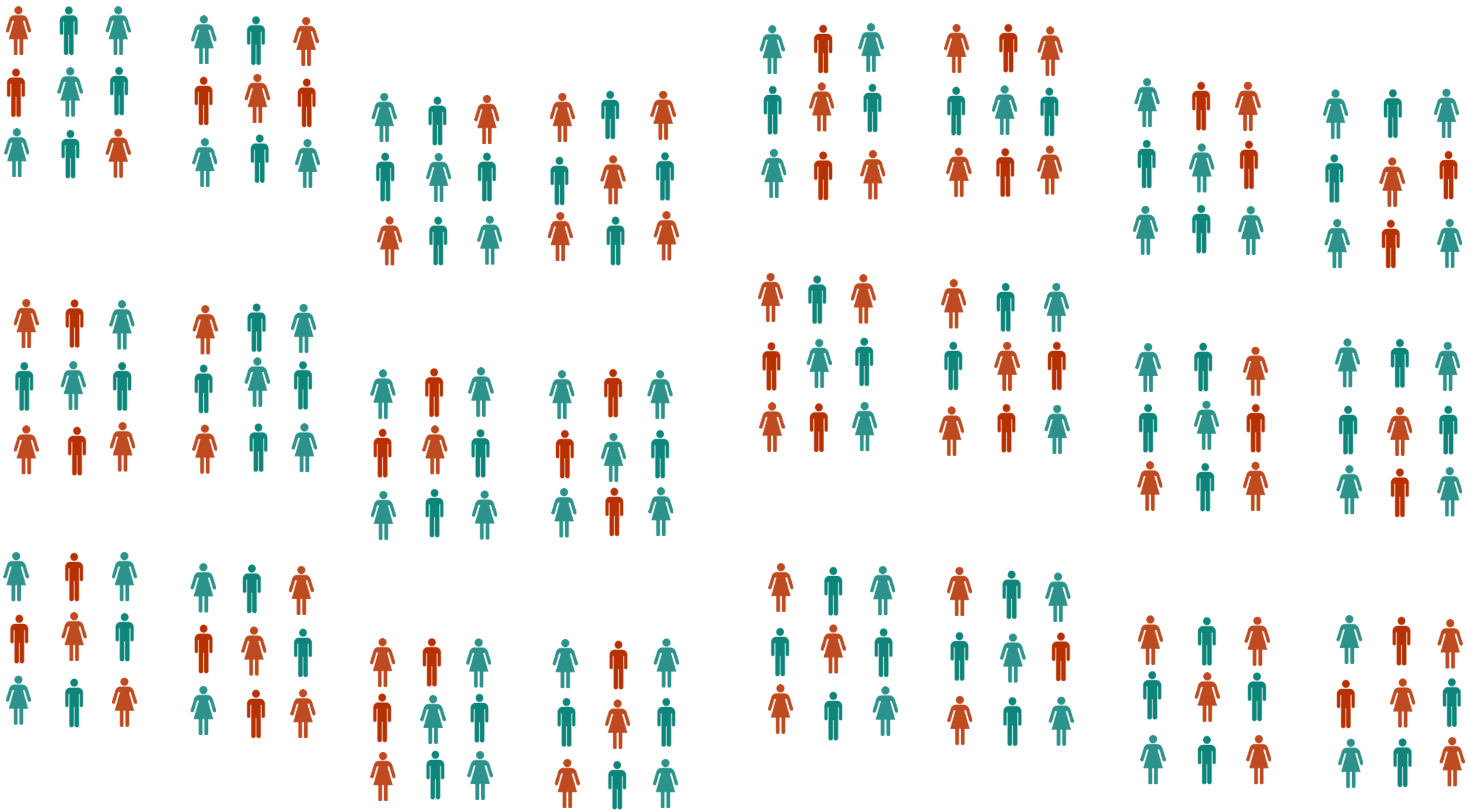
- Introduction
- Hypothesis testing
- What influences power?
  - Effect size
  - Sample size
  - Variance
  - Proportion of sample in treatment vs. control
- **Power in clustered designs**
- Calculating power in practice

Power in clustered designs

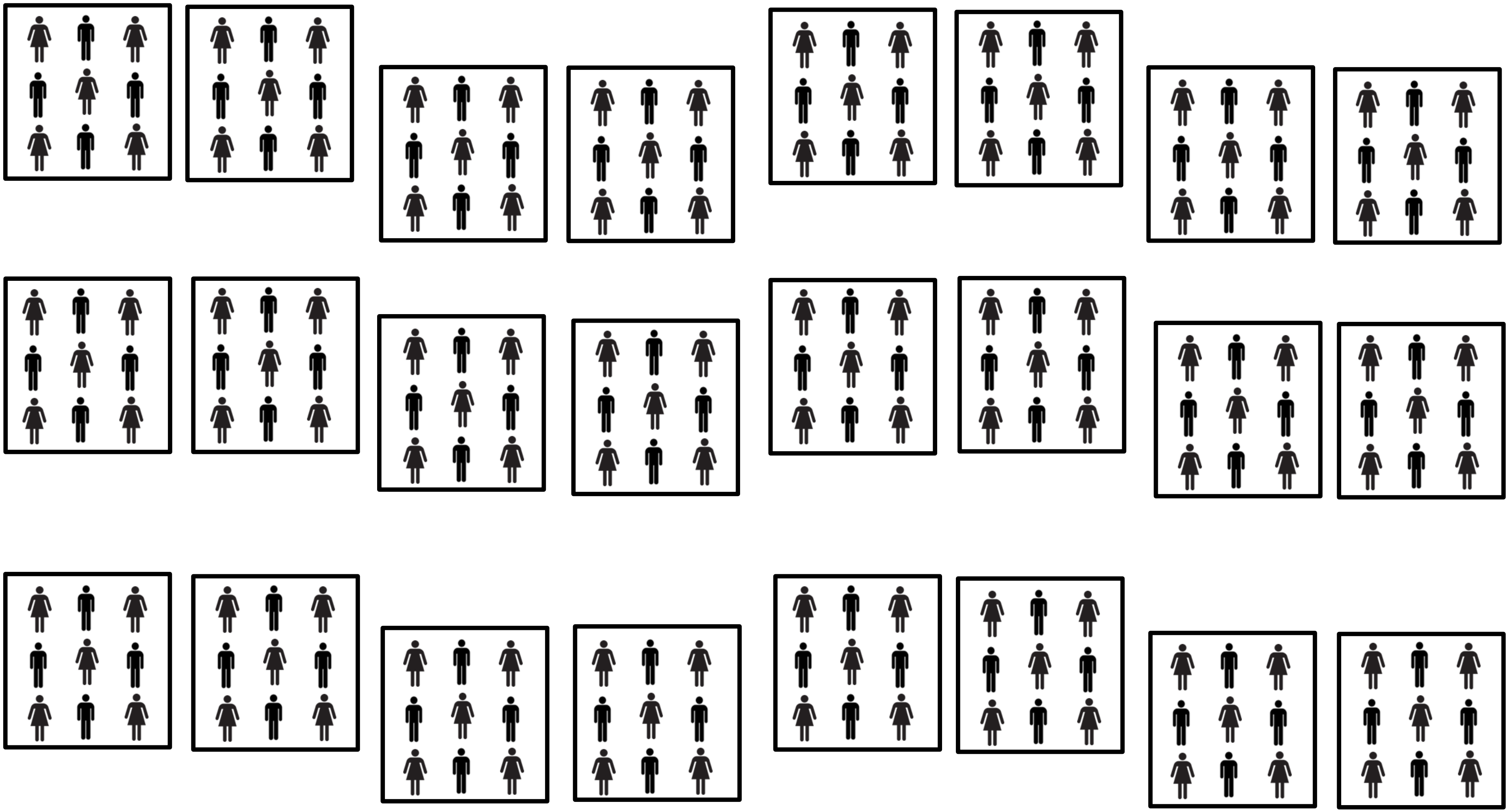
Randomize individuals to T or C



Randomize individuals to T or C



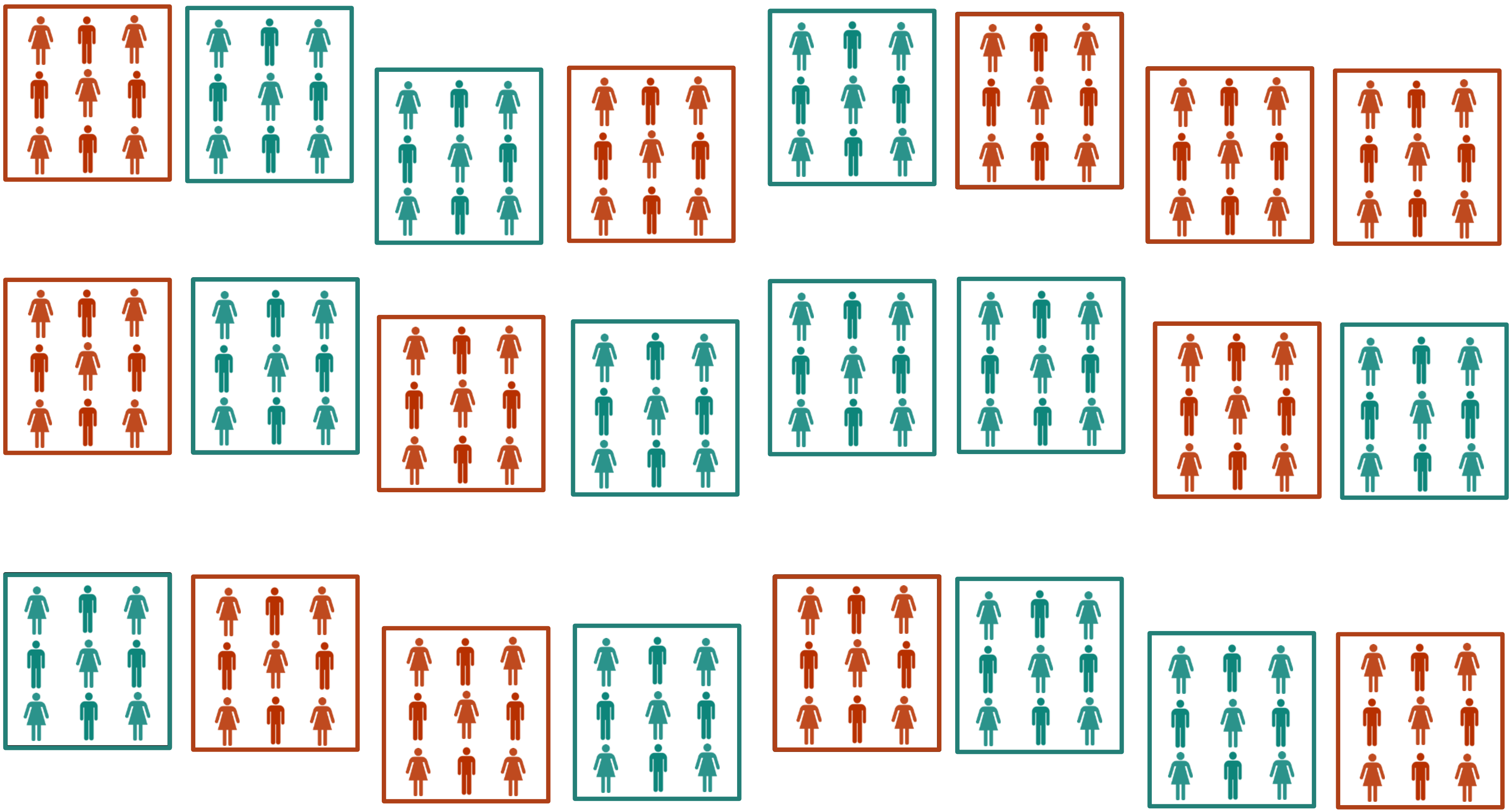
Or randomize clusters: e.g. sports teams



…like cricket

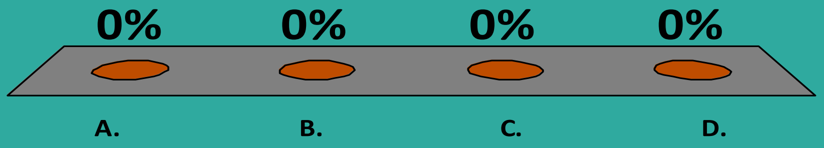
- Each team assigned to Individual Pay vs. Team Pay
- How does this affect social cohesion?

Or randomize clusters: e.g. cricket teams



Compared to an individual level randomized design, to achieve the same power, a clustered level RCT is likely to require…

A. A smaller sample size
B. A bigger sample size
C. The same sample size
D. Don't know



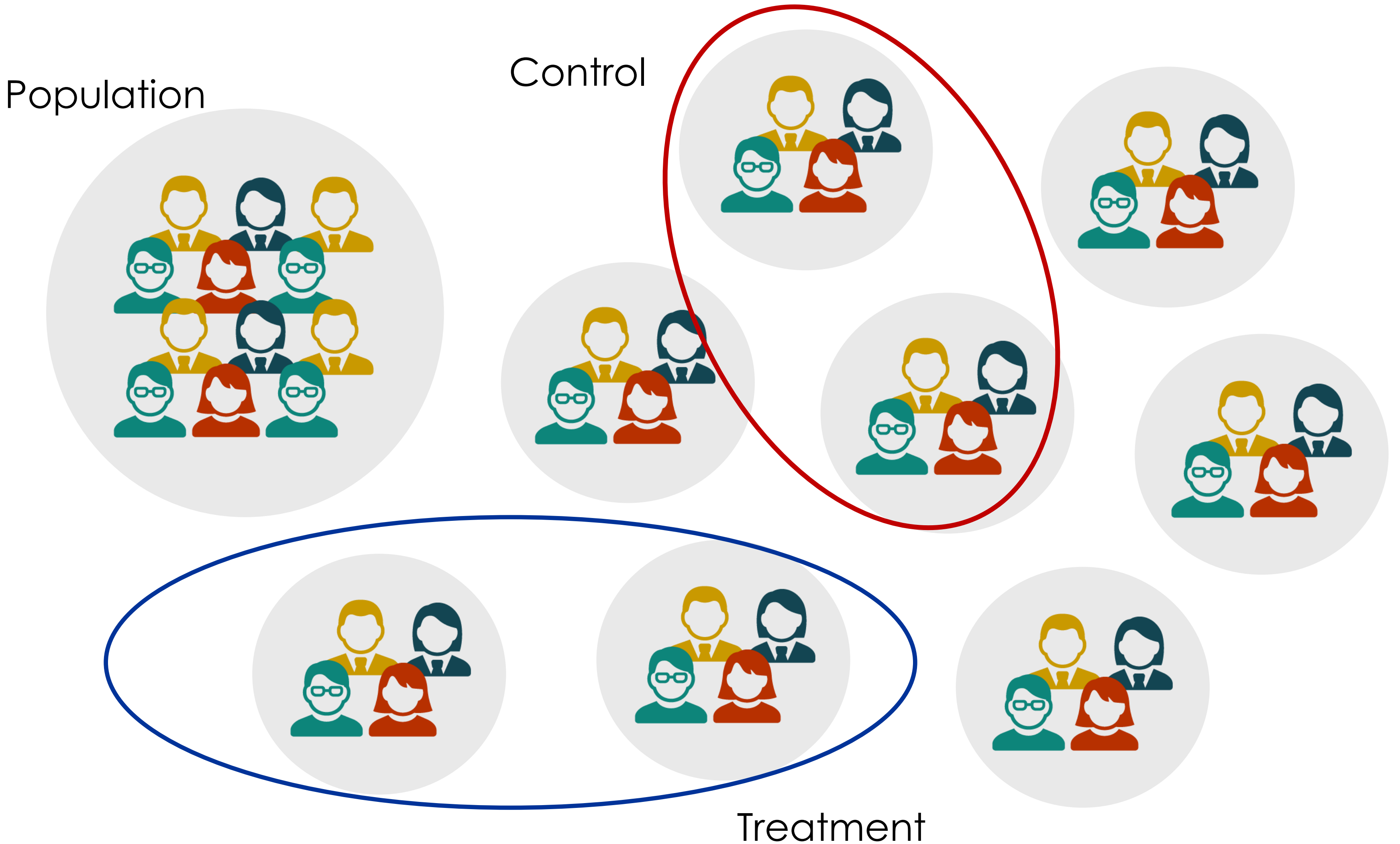
Clustered design: intuition

- You want to know how close the upcoming Lok Sabha elections will be

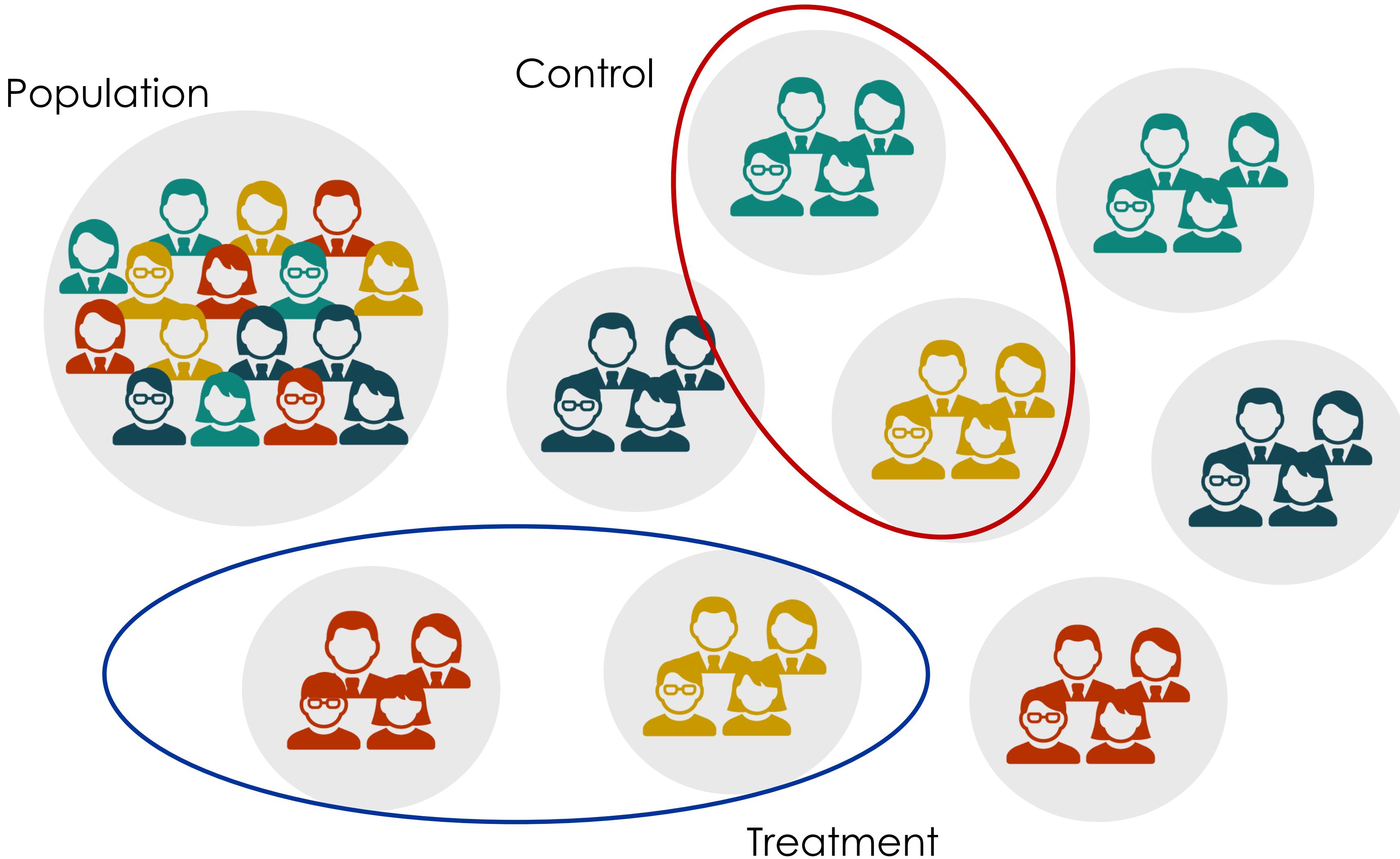
- Method 1: Randomly select 50 people from entire Indian population

- Method 2: Randomly select 10 families, and ask five members of each family their opinion

Low intra-cluster correlation (ICC) or ρ (rho)



HIGH intra-cluster correlation (ρ)



Intra-cluster correlation definition

- Total variance can be divided into cluster-level variance (τ^2) and remaining variance (σ^2)
- When variance within clusters is small, the within cluster correlation is high, i.e., (ICC) is high (previous slide)
- Definition of ICC: the proportion of total variation explained by cluster-level variance

$$icc = \rho = \frac{\tau^2}{\sigma^2 + \tau^2}$$

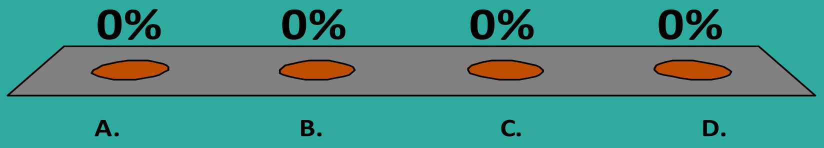
- Simply: the ICC describes how strongly units in the same group resemble each other

How does ICC impact power?

- For a given N we have less power when we randomize by cluster (unless ICC is zero)
- There are diminishing returns to surveying more people per cluster
- Usually the number of clusters is the key determinant of power, not the number of people per cluster

All uneducated people live in one village. People with only primary education live in another. College grads live in a third, etc. ICC (ρ) on education will be..

A. High
B. Low
C. No effect on rho
D. Don’t know



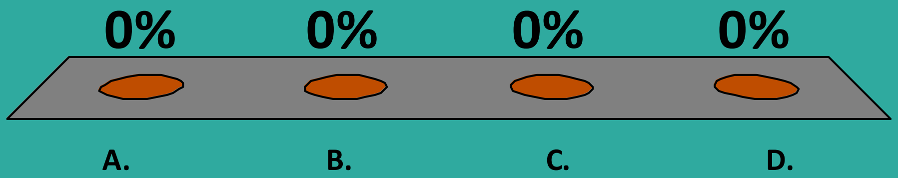
If ICC (ρ) is high, what is a more efficient way of increasing power?

A. Include more clusters in the sample

B. Interview more people in each cluster

C. Both

D. Don't know



Power with clustering

Significance Level

Effect Size

Power

Variance

$$\frac{EffectSize}{\sqrt{1+\rho(m-1)}} = \left(t_{(1-\kappa)} + t_{\alpha}\right) * \sqrt{\frac{1}{P(1-P)}} * \sqrt{\frac{\sigma^2}{N}}$$

ICC

Average Cluster Size

Proportion in Treatment

Sample Size

Outline

- Introduction
- Hypothesis testing
- What influences power?
  - Effect size
  - Sample size
  - Variance
  - Proportion of sample in treatment vs. control
- Power in clustered designs
- **Calculating power in practice**

Calculating power in practice

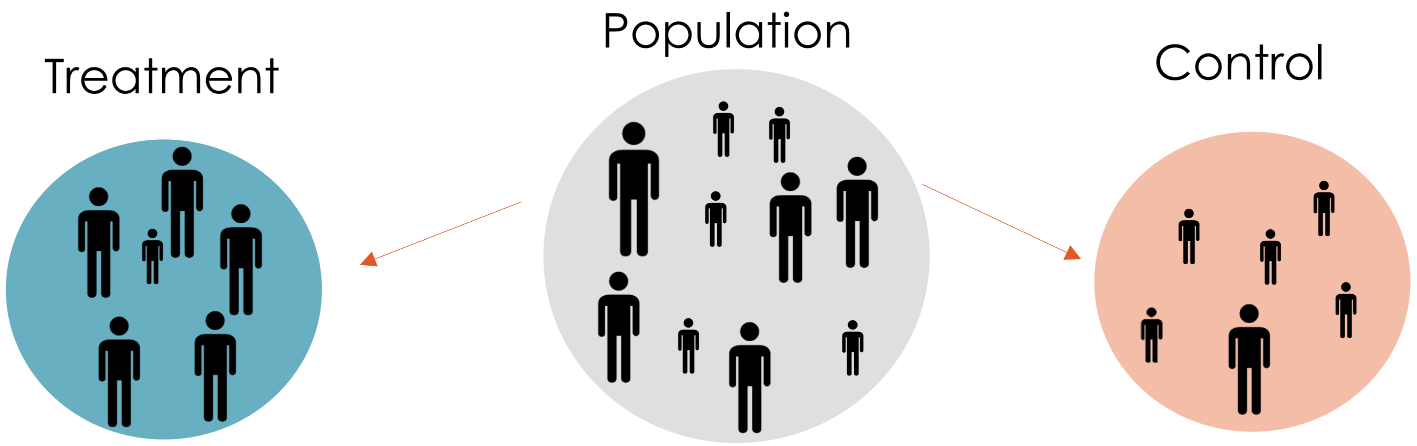
Finding the ingredients for the power equation

Calculating power: A step-by-step guide

1. Set desired power (80%, 90%) and significance (95%)
2. Calculate residual variance (& ICC) using pilot data, national data sources, or data from other studies
3. Decide number of treatments
4. Set MDE size for T vs C and between treatments
5. Decide allocation ratio
6. Calculate sample size
7. Estimate resulting budget
8. Adjust parameters above (e.g. cut number of arms)
9. Repeat

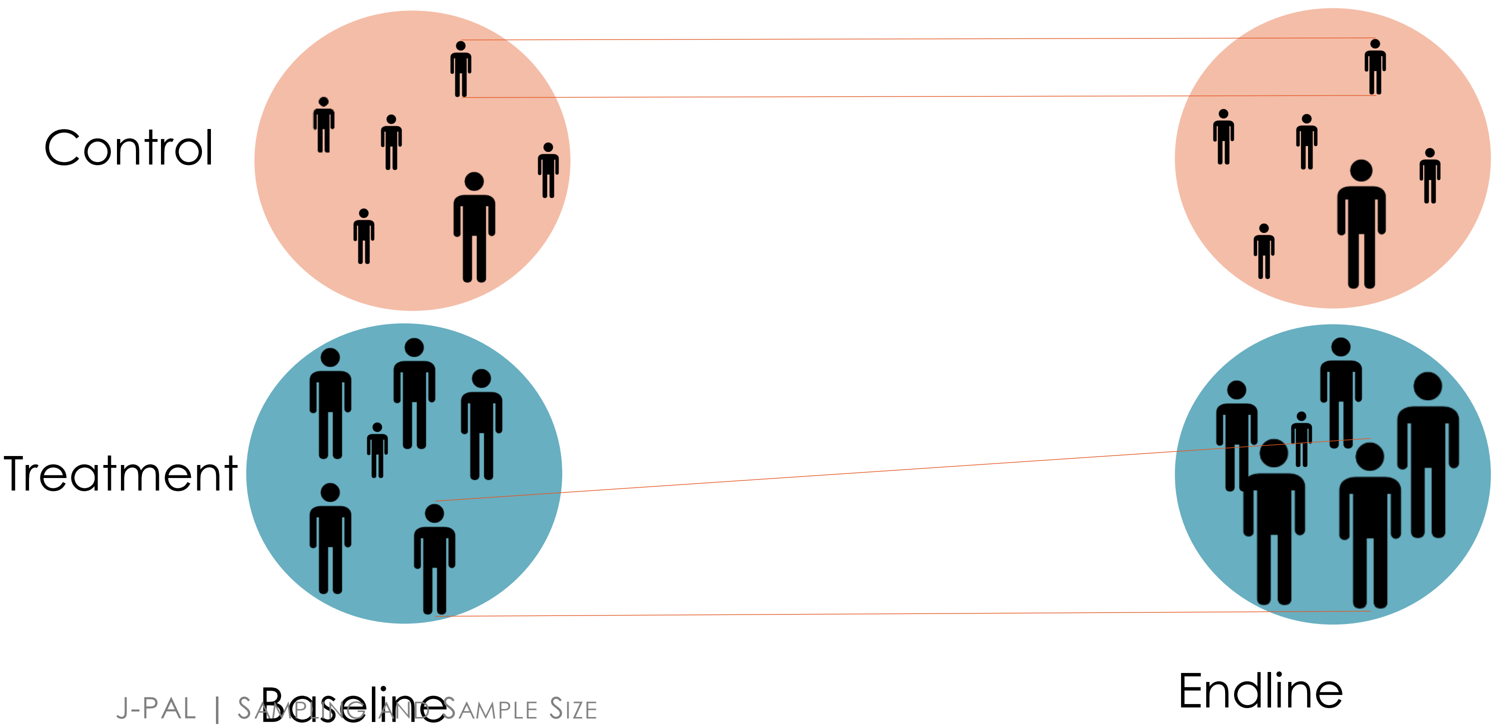
Variance

- Estimate variance from data from similar populations
  - Not much we can do about population variance
- Can help distinguish if endline height difference due to:
  - Program increasing height
  - Chance allocation of tall children to treatment at baseline



Residual variance: intuition

- Some part of endline height can be explained by baseline height.
- Accounting for this allows a more *precise* estimate of intervention effect.



Residual variance

- Variance reduces power because of the risk that we might pick "successful" people for treatment
- Some variation can be explained by observables
  - Older kids are taller
- Using controls in analysis soaks up variance, impact more precisely estimated, more power
- Calculate residual variance by regressing outcome on controls in existing data
- Baseline value of outcome good control
  - In Stata power, can adjust for multiple rounds of data
  - Need estimate of correlation in outcome between rounds

Estimating rho

- Rho must be between 0 and 1
- Depends on context and outcome variable
- To estimate Rho need big samples
- Check sensitivity of your power to different possible rhos

Estimating rho

Malawi: Households produces maize	0.003
Sierra Leone: Households produce cocoa	0.57
Sierra Leone: Average rice yields	0.04
Busia, Kenya: Math and language test scores	0.22
Busia, Kenya: Math test scores	0.62
Mumbai, India: Math and language test scores	0.28
Texas: Precinct voting preferences	0.20
Italy: Hospital admissions	0.06
United States: Weight and cholesterol levels	0.02
United States: Reading achievement scores	0.22

Number of treatment arms

- Different treatment arms help disentangle different mechanisms behind an effect
- More arms requires lots more sample size
- Have to worry about two different MDEs
  - for comparing T vs C
  - Comparing T_1 vs T_2

Multiple arms: Bangladesh example

	Age group	BDHS 2007	BDHS 2011
% Completed secondary education	15-19	9	19
% married before age 18	20-24	66	65
% begun childbearing	15-19	33	30
% used contraception	15-19	42	47
% gave birth at home	<20	86	71

Kishore Kontha (KK) in Bangladesh

Does basic KK work?

(1)	Basic	Gender attitudes, literacy education, health information	76
(2)	Livelihoods	Basic plus financial literacy (half, 39, receive savings club)	77
(3)	Full	Livelihoods plus oil (half, 38, receive savings club)	77
(4)	Oil	Oil incentive for unmarried girls	77
(5)	Control		153
Total			460
(6)	Savings cross cut	Savings club cross cut with Full and Livelihoods	77

Kishore Kontha (KK) in Bangladesh

Marginal impact of fin lit

(1)	Basic	Gender attitudes, literacy education, health information	76
(2)	Livelihoods	Basic plus financial literacy (half, 39, receive savings club)	77
(3)	Full	Livelihoods plus oil (half, 38, receive savings club)	77
(4)	Oil	Oil incentive for unmarried girls	77
(5)	Control		153
Total			460
(6)	Savings cross cut	Savings club cross cut with Full and Livelihoods	77

Kishore Kontha (KK) in Bangladesh

Did KK work?

(1)	Basic	Gender attitudes, literacy education, health information	76
(2)	Livelihoods	Basic plus financial literacy (half, 39, receive savings club)	77
(3)	Full	Livelihoods plus oil (half, 38, receive savings club)	77
(4)	Oil	Oil incentive for unmarried girls	77
(5)	Control		153
Total			460
(6)	Savings cross cut	Savings club cross cut with Full and Livelihoods	77

153

153

Kishore Kontha (KK) in Bangladesh

Impact of delayed marriage incentive

(1)	Basic	Gender attitudes, literacy education, health information	76
(2)	Livelihoods	Basic plus financial literacy (half, 39, receive savings club)	77
(3)	Full	Livelihoods plus oil (half, 38, receive savings club)	77
(4)	Oil	Oil incentive for unmarried girls	77
(5)	Control		153
Total			460
(6)	Savings cross cut	Savings club cross cut with Full and Livelihoods	77

154

153

Kishore Kontha (KK) in Bangladesh

(1)	Basic	Gender attitudes, literacy education, health information	76
(2)	Livelihoods	Basic plus financial literacy (half, 39, receive savings club)	77
(3)	Full	Livelihoods plus oil (half, 38, receive savings club)	77
(4)	Oil	Oil incentive for unmarried girls	77
(5)	Control		153
Total			460
(6)	Savings cross cut	Savings club cross cut with Full and Livelihoods	77

Impact of savings clubs

	Saving	Not saving
(2) →	38	39
(3) →	39	38
	77	77

77 vs 77

Multiple treatment arm tips

- Good to have at least one intensive arm, where a zero would be interesting
- In the analysis, will it be useful to pool all the treatment arms, creating an "any treatment" arm
- Have to scale up sample size given from statistical packages
- Stata gives sample per cell
  - 2 treatments + C = 3 cells
- Optimal Design assumes one treatment, one control
  - Divide result by 2 to get sample size per cell

Unequal allocation ratio

- If budget covers treatment and evaluation, more expensive to add one person to T than C
  - Unequal allocation ratio gives a bigger total N, which may be worth it
  - Try different allocation ratios within a given budget and explore tradeoff
- With multiple treatments put more sample behind most important question
  - If going to pool treatments, have bigger control
  - If really care about between treatments need bigger sample in treatment groups

Unequal allocation: example

(1)	Basic	Gender attitudes, literacy education, health information	76
(2)	Livelihoods	Basic plus financial literacy (half, 39, receive savings club)	77
(3)	Full	Livelihoods plus oil (half, 38, receive savings club)	77
(4)	Oil	Oil incentive for unmarried girls	77
(5)	Control		153
Total			460
(6)	Savings cross cut	Savings club cross cut with Full and Livelihoods	77

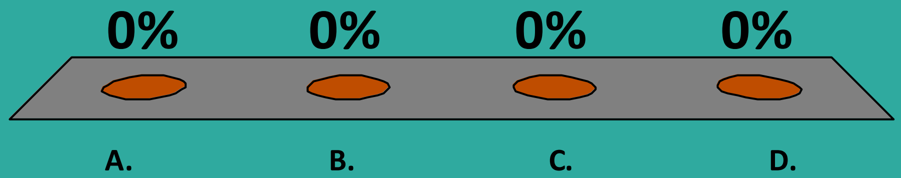
What effect size should you use when designing your experiment?

A. Smallest effect size that is still cost effective

B. Largest effect size you expect your program to produce

C. Both

D. Neither



Minimum detectable effect size

- The most important ingredient for calculating power
- MDE is not the effect size we expect or want
- MDE is the effect size below which we may not be able to distinguish the effect from zero, even if it exists
  - i.e. below which effect might as well be zero

Questions to ask when determining MDE

- Below what effect size would the program not be cost effective
- How big an effect would we need to make the result interesting?
- May want smaller MDE between arms than between T and C
- Common mistake is powering on T vs. C so don't have power to distinguish between arms

	Units	Obs/ village	MDE v=77	MDE v=154	Mean	SD	ICC
Age at Marriage	yrs	42	0.13	0.09	18.2	2.63	0.03
Reproductive Health							
Age at first birth	yrs	42	0.13	0.09	20.0	2.1	0.03
Number of births	count	42	0.14	0.10	2.0	0.7	0.04
Number of antenatal visit	count	6	0.25	0.18	3.0	3.3	0.04
Pregnancy complications	ppts	42	-8.57	-6.15	75.0	43.6	0.00
Physical Health							
Mothers' weight	kg	6	0.26	0.19	49.3	9.4	0.07
Arm circumference	cm	6	0.23	0.17	26.2	3.6	0.01
Hemoglobin	gm/dL	6	0.26	0.18	11.9	1.1	0.03
Mental Health							
Cortisol	pg/mg	6	-0.27	-0.19	5.1	3.1	0.08
Depression per week	days	6	-0.26	-0.18	2.1	1.1	0.05
Child Health							
Breastfeeding	months	6	0.27	0.19	15.3	9.8	0.08
Children's weight for age	kg/yr	6	0.23	0.17	7.3	7.8	0.01
--, gender difference	kg/yr	6	0.26	0.18	0.7	7.8	0.05
Knowledge							
Risks to young mothers	count	6	0.27	0.19	2.5	1.3	0.07
Decision-Making							
Independent decisions	count	6	0.25	0.18	2.1	1.4	0.04
Attitudes & Empowerment							
Earliest marriage age	yrs	6	0.23	0.16	18.4	1.6	0.00
Education							
Last class passed	yrs	42	0.20	0.14	9.1	4.2	0.11

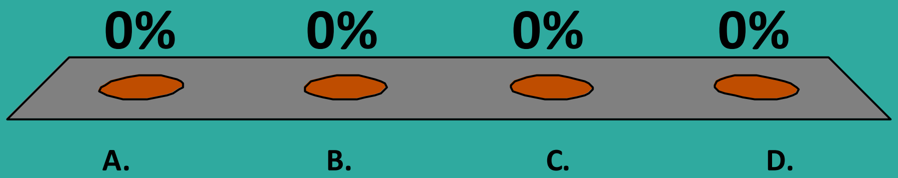
Which statement is NOT correct?

A. In general, cheaper programs can be tested with smaller sample sizes

B. When calculating MDE we need to think about likely take up rates

C. A smaller MDE means the effect will be estimated more precisely (all else equal)

D. MDE has a big influence on power, for a given sample size



Calculating power in Stata

- Stata has a new command “power” where you can state sample size and get out power
  - But does not allow for clustering
- Most still use sampsi and sampclus (add ons) or clustersampsi
  - Default is power 90%, significance 5%, equal allocation
- To detect an increase in average test scores from 43% to 45% with power of 80%:

  sampsi 0.43 0.45, power(0.8) sd(0.05)

- Stata gives N per cell e.g. N1=99
  - With multiple arms, need to multiply by number of cells (i.e. number of treatments plus control)
- For binary outcomes, SD determined by mean

Power in Stata with clustering

- First calculate sample size without clustering and then add information on clusters

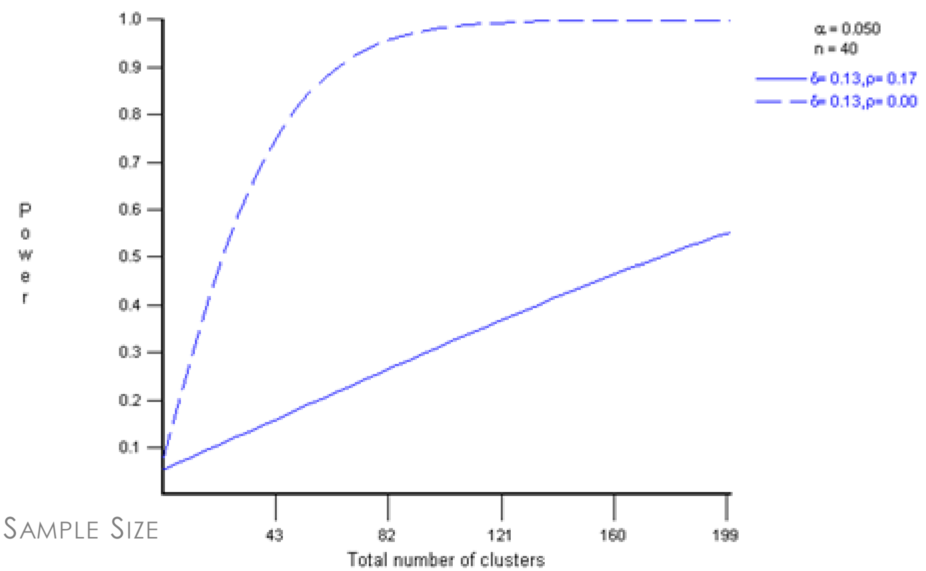
- If above experiment were to be randomized at class level with 60 per class and ICC of 0.2

```
sampsi 0.43 0.45, power(0.8) sd(0.05)

sampclus, obsclus(60) rho(0.2)
```

Power with optimal design

- Optimal design is a free software specifically designed for power calculations
- MDE must be entered in standardized effect size (i.e. effect size divided by standard deviation)
- OD allows multiple levels of clustering and works with dropdown menus—see JPAL exercise for details



END!

Follow us on Twitter @JPAL_SA
Visit our website www.povertyactionlab.org for more resources