



Planning sample size for randomized evaluations

Simone Schaner

Dartmouth College

Course Overview

- Why evaluate? What is evaluation?
- Outcomes, indicators and measuring impact
- Impact evaluation – why randomize
- How to randomize
- **Sampling and sample size**
- Implementing an evaluation
- Analysis and inference

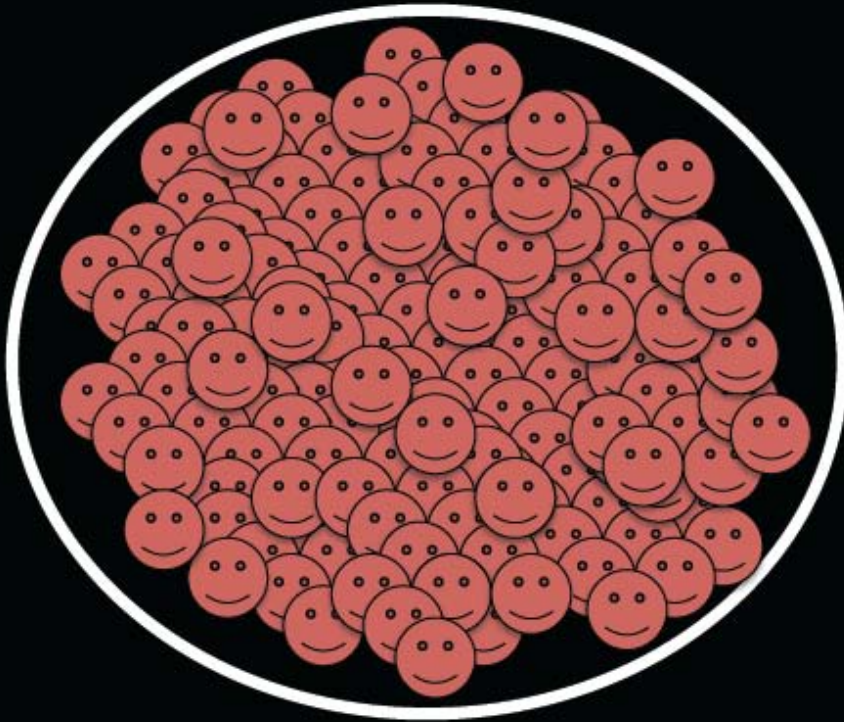
Today's Question

- How large does the sample need to be to “credibly” detect a given treatment effect?
- What does credibly mean?
- Randomization removes bias, but it does not remove noise
- But how large must “large” be?

Lecture Overview

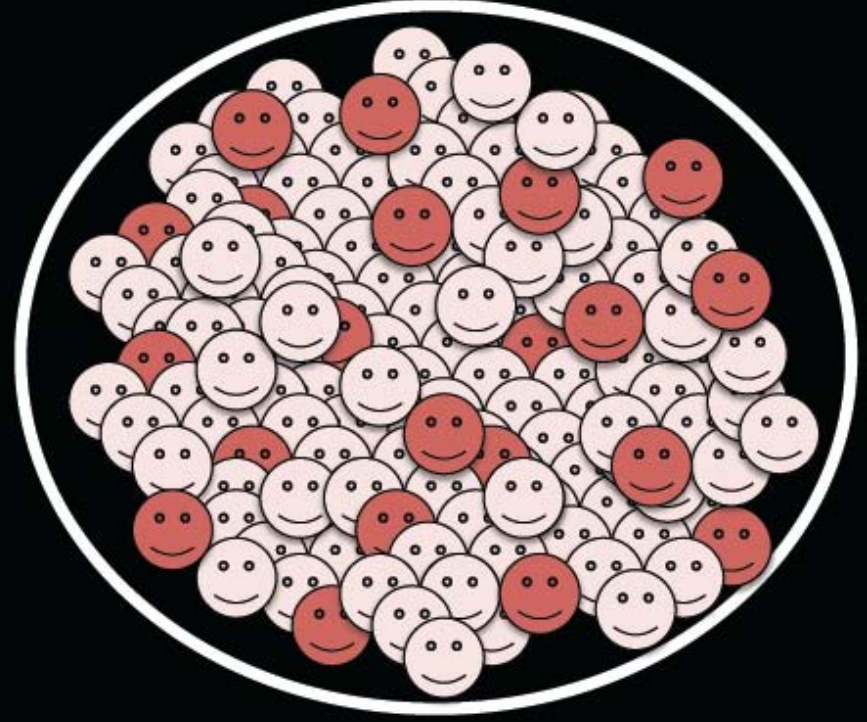
- Estimation
- Intro to the scientific method
- Hypothesis testing
- Statistical significance
- Factors that influence power
- Effect size
- Sample size
- Cluster randomized trials

Estimation



Population

We wish to learn about this

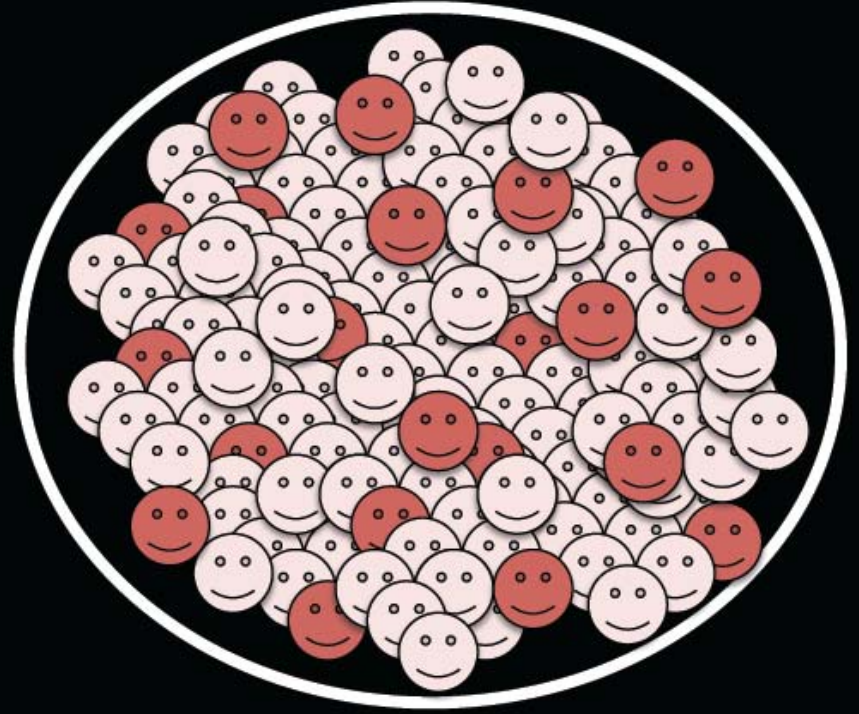
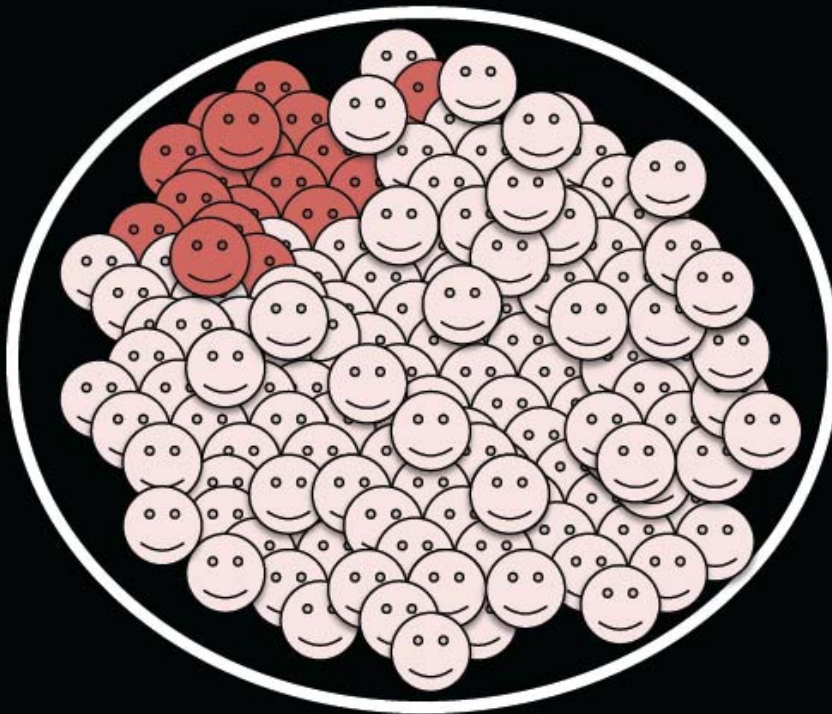


Sample

But we only see this

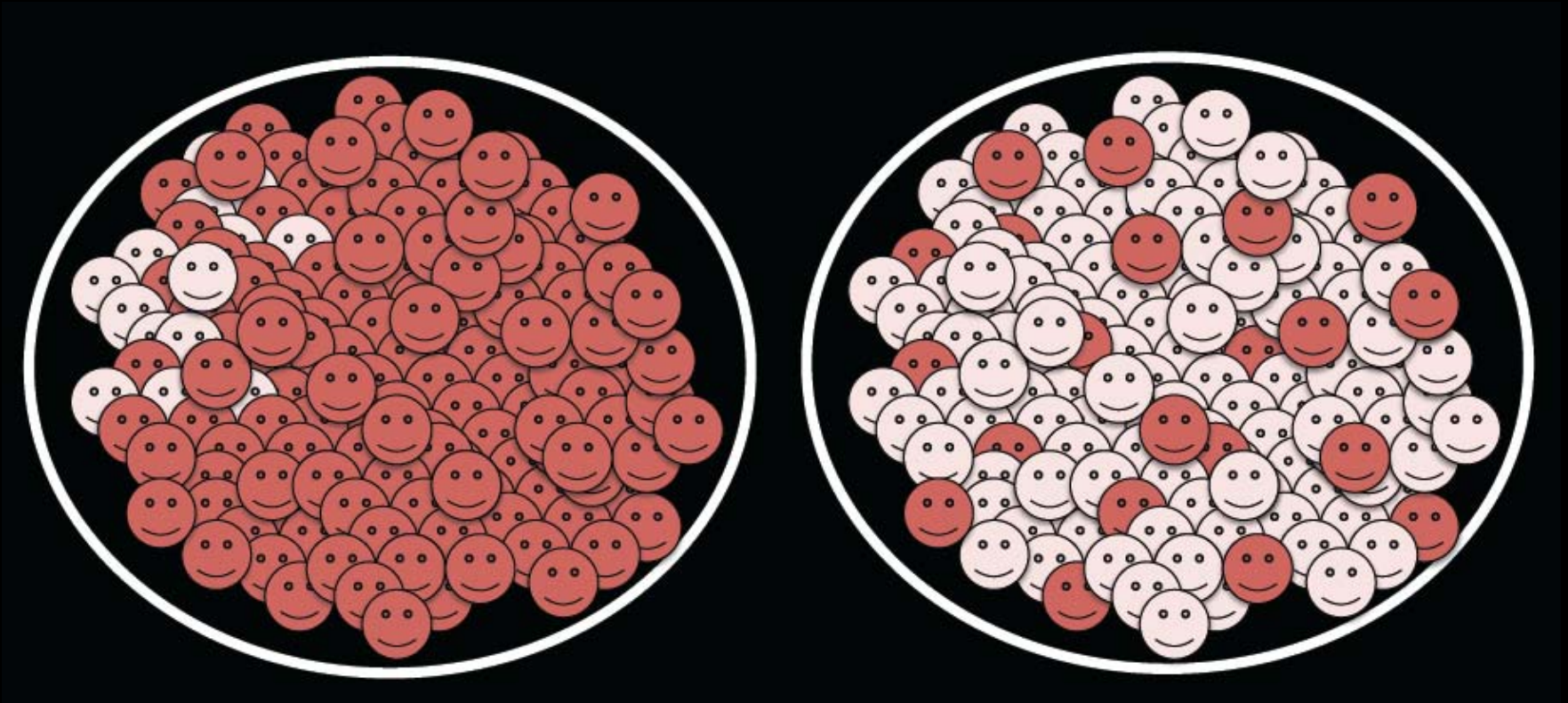
The *sample average* is our estimate of the *population average*

Accuracy: Estimate is Right *On Average*



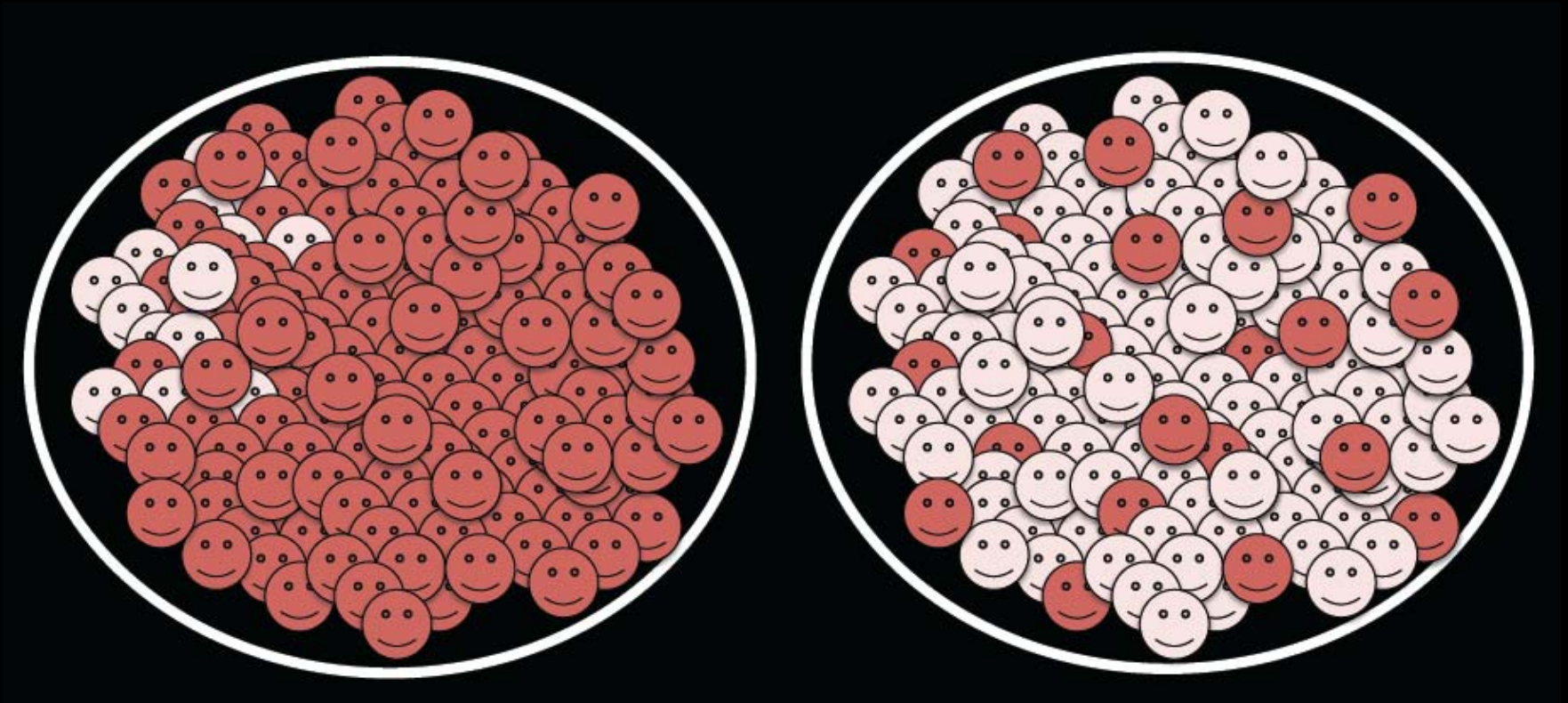
Which sampling strategy will give us a more accurate estimate?

Precision: Estimate Has Low Variability



Which sampling strategy will give us a more precise estimate?

Precision: Estimate Has Low Variability



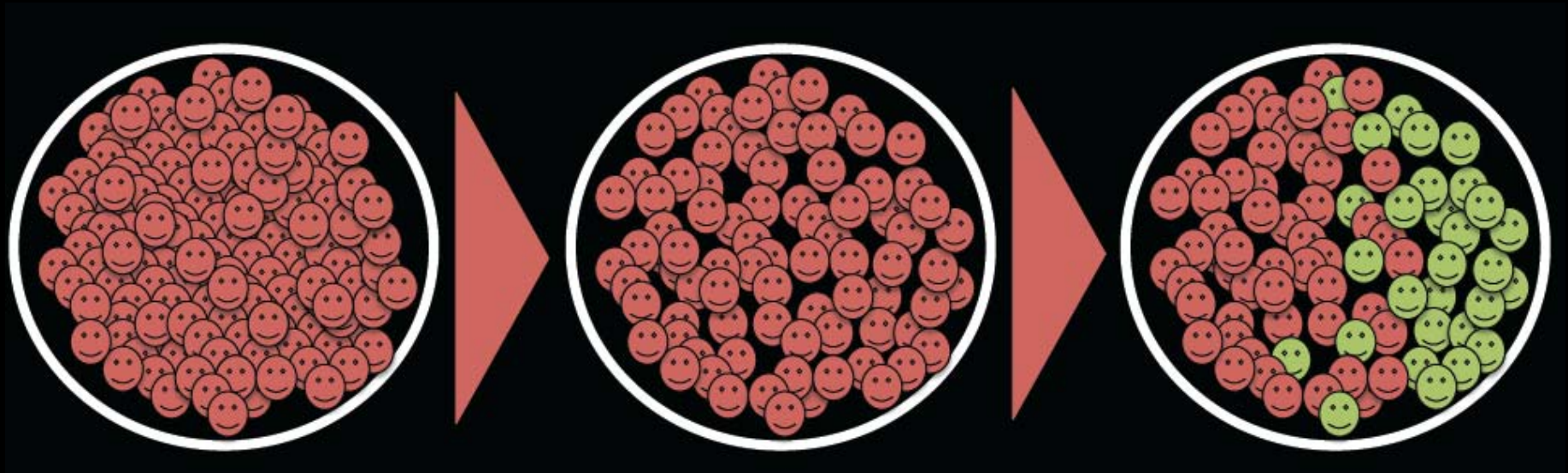
But what about a more *accurate* estimate?

Estimation

- When we do estimation
- Sample size allows us to say something about the variability of our estimate
- But it doesn't ensure that our estimate will be close to the truth on average

RANDOMIZATION IS THE GOLD STANDARD BECAUSE IT ENSURES ACCURACY. We then control precision with sample size.

Review: Random Sampling vs. Random Assignment to Treatment

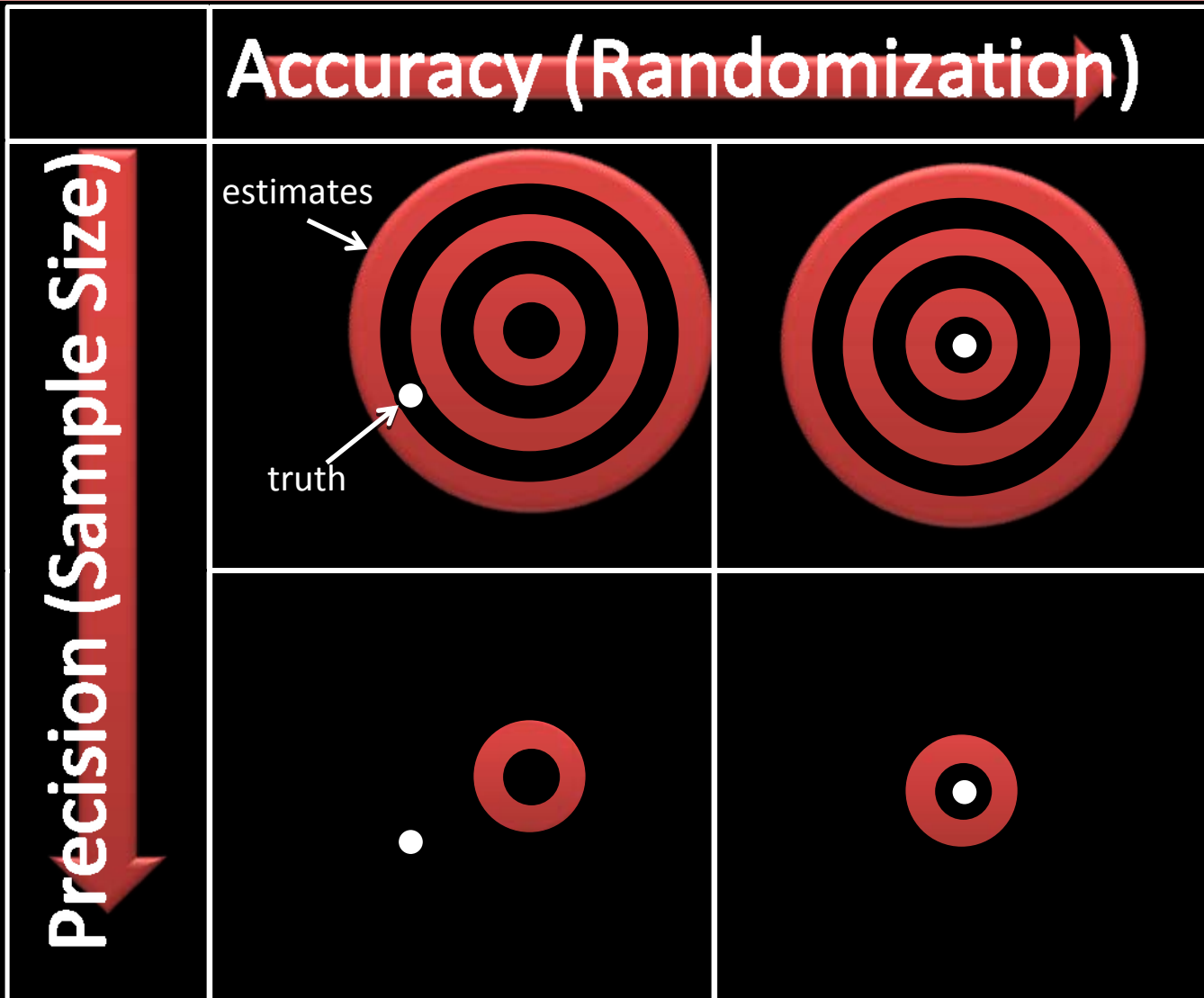


What happens if we
randomly sample...

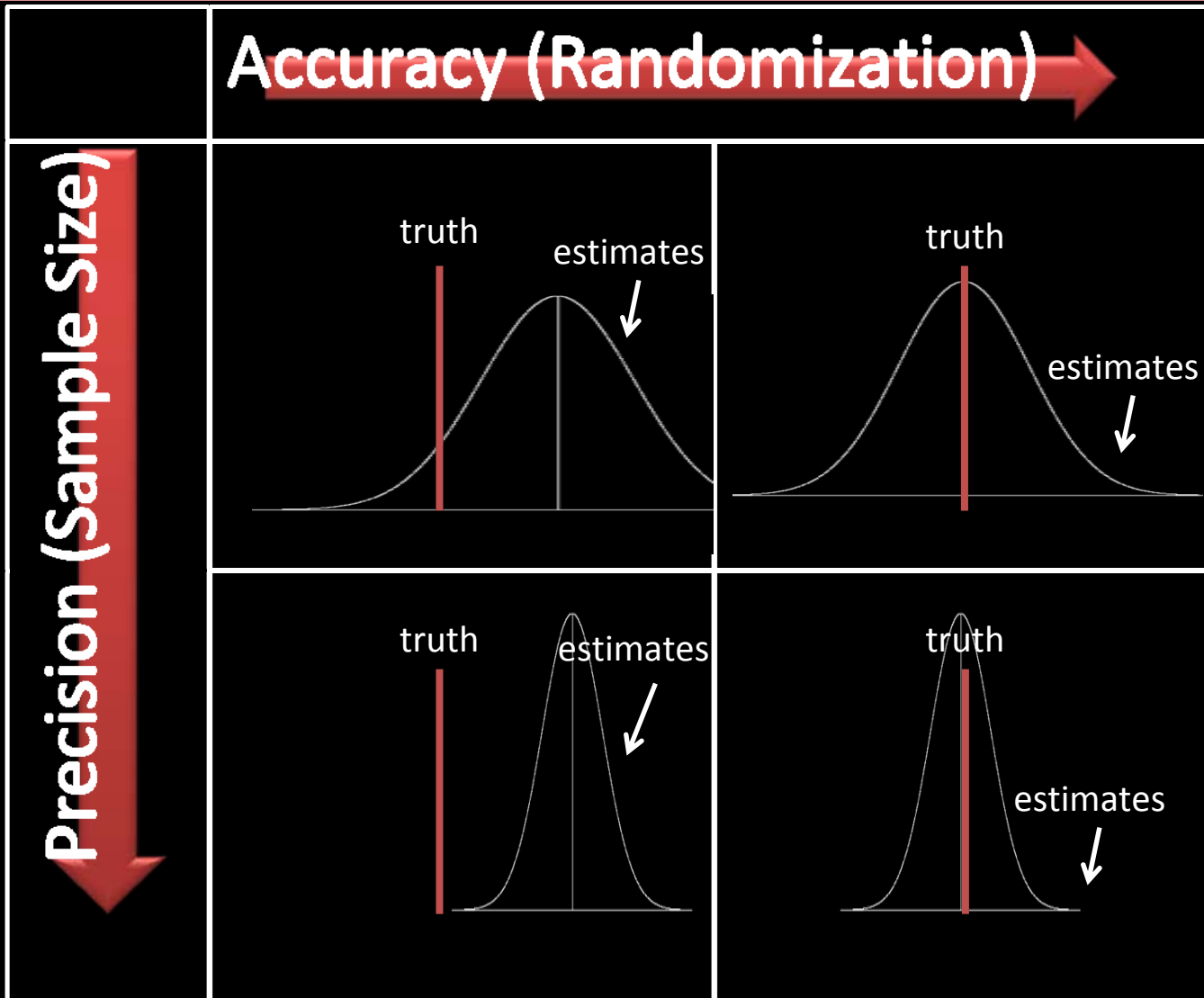
...But don't randomly
assign treatment?

Will our estimate of the treatment effect be unbiased?

Accuracy versus Precision



Accuracy versus Precision



Measuring Significance: Scientific Method

- Does the scientific method apply to social science?
- The scientific method involves:
 - 1) proposing a hypothesis
 - 2) designing experimental studies to test the hypothesis
- How do we test hypotheses?

Basic set up

- We start with our hypothesis
- At the end of an experiment, we test our hypothesis
- We compare the outcome of interest in the treatment and the comparison groups.

Hypothesis testing

- In criminal law, most institutions follow the rule: “innocent until proven guilty”
- The prosecutor wants to prove their hypothesis that the accused person is guilty
- The burden is on the prosecutor to show guilt
- The jury or judge starts with the “null hypothesis” that the accused person is innocent

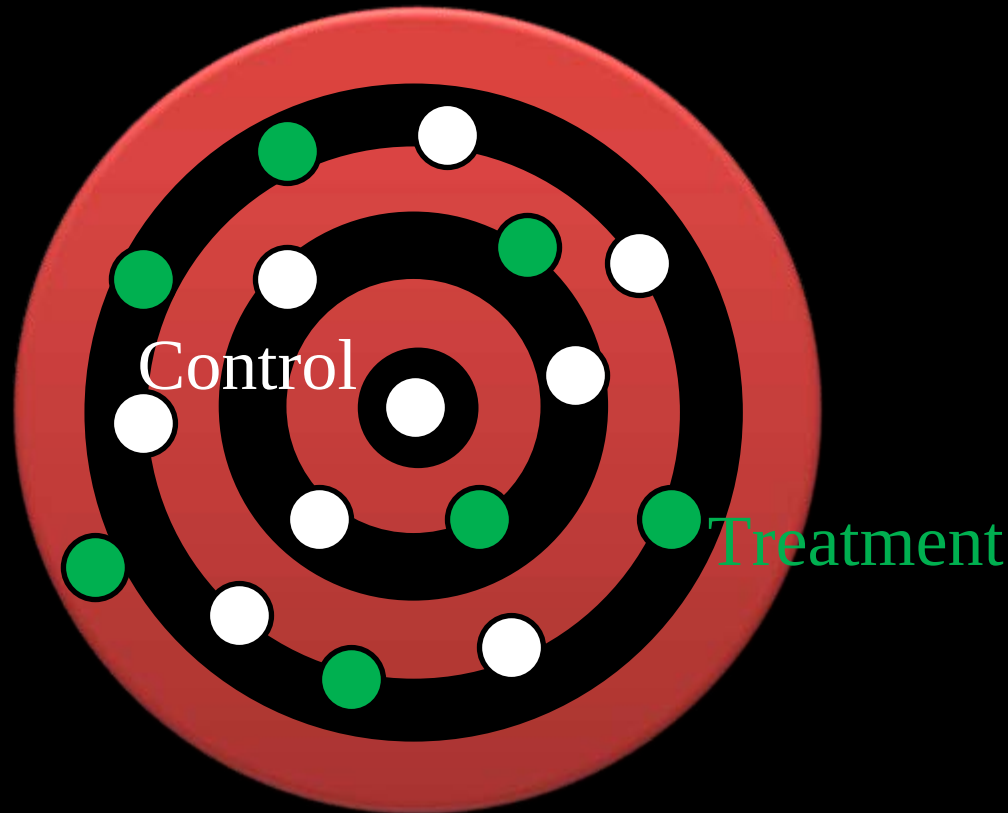
Hypothesis testing

- In program evaluation, instead of “presumption of innocence,” the rule is: “presumption of insignificance”
- Policymaker’s hypothesis: the program improves learning
- Evaluators approach experiments using the hypothesis:
 - “There is zero impact” of this program
 - Then we test this “Null Hypothesis” (H_0)
- The burden of proof is on the program
 - Must show a statistically significant impact

Hypothesis testing

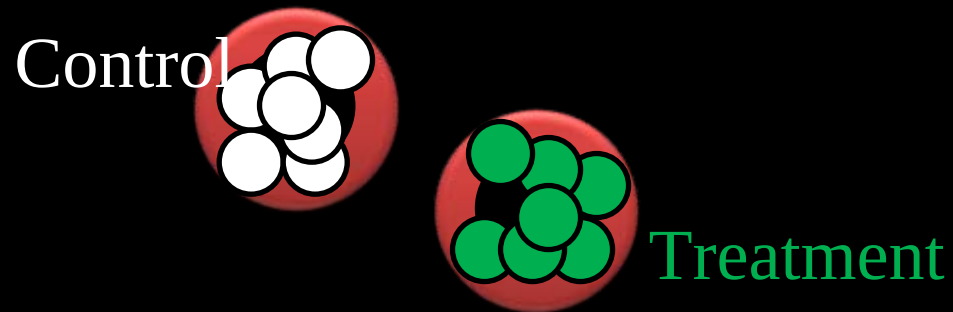
- If our measurements show a difference between the treatment and control group, our first assumption is:
 - In truth, there is no impact (our H_0 is still true)
 - There is some margin of error due to sampling
 - “This difference is solely the result of chance (random sampling error)”
- We (still assuming H_0 is true) then use statistics to calculate how likely this difference is in fact due to random chance

Is this difference due to random chance?



Probably...

Is this difference due to random chance?



Probably not....

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”

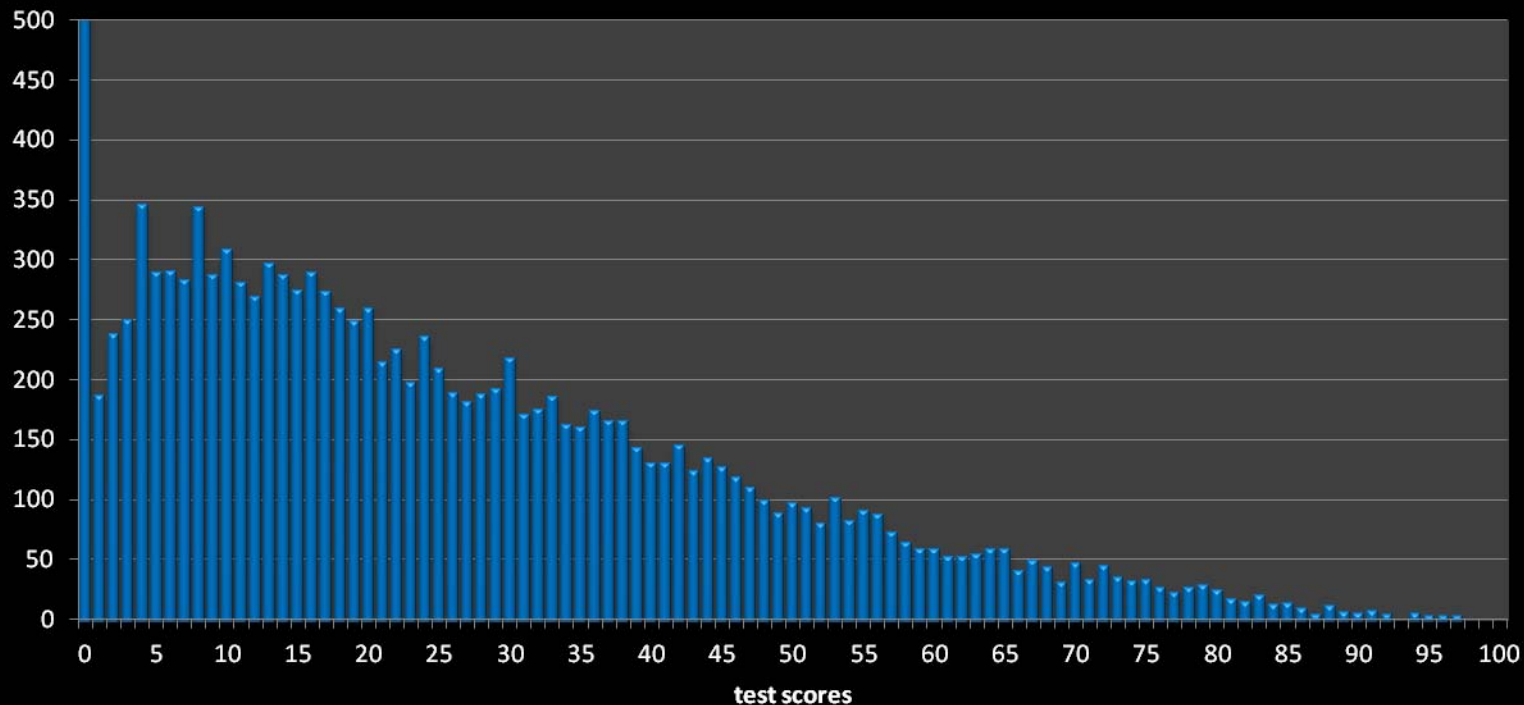
Hypothesis testing: conclusions

- Are we now 100 percent certain there is an impact?
 - No, we may be only 95% confident
 - And we accept that if we use that 5% threshold, this conclusion may be wrong 5% of the time
 - That is the price we're willing to pay since we can never be 100% certain
 - **Because we can never see the counterfactual, We must use random sampling and random assignment, and rely on statistical probabilities**

Example: Pratham Balsakhi (Vadodarda)

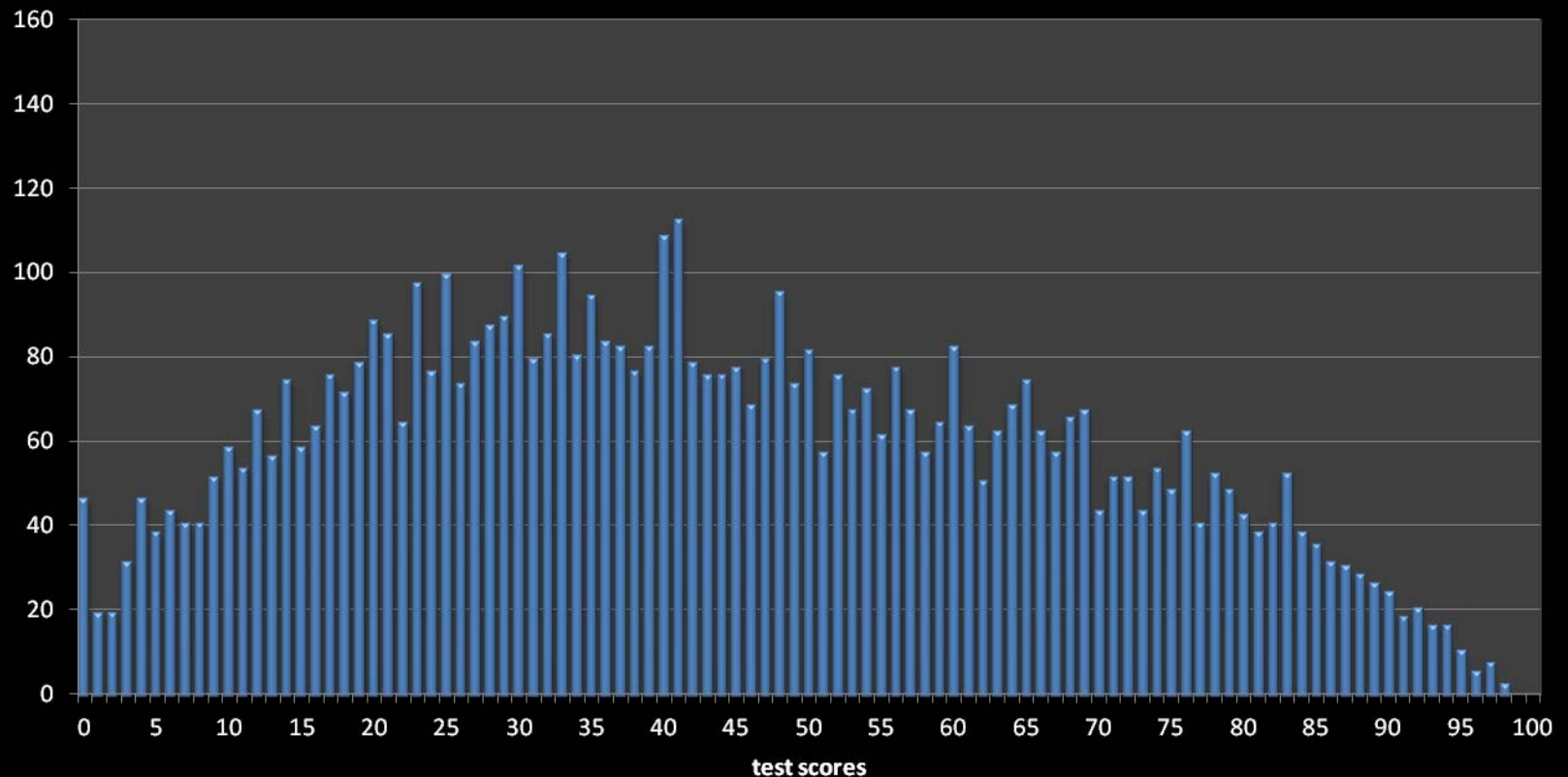


Baseline test score data in Vadodara



- This was the distribution of test scores in the baseline.
- The test was out of 100.
- Some students did really well, most, not so well
- Many actually scored zero

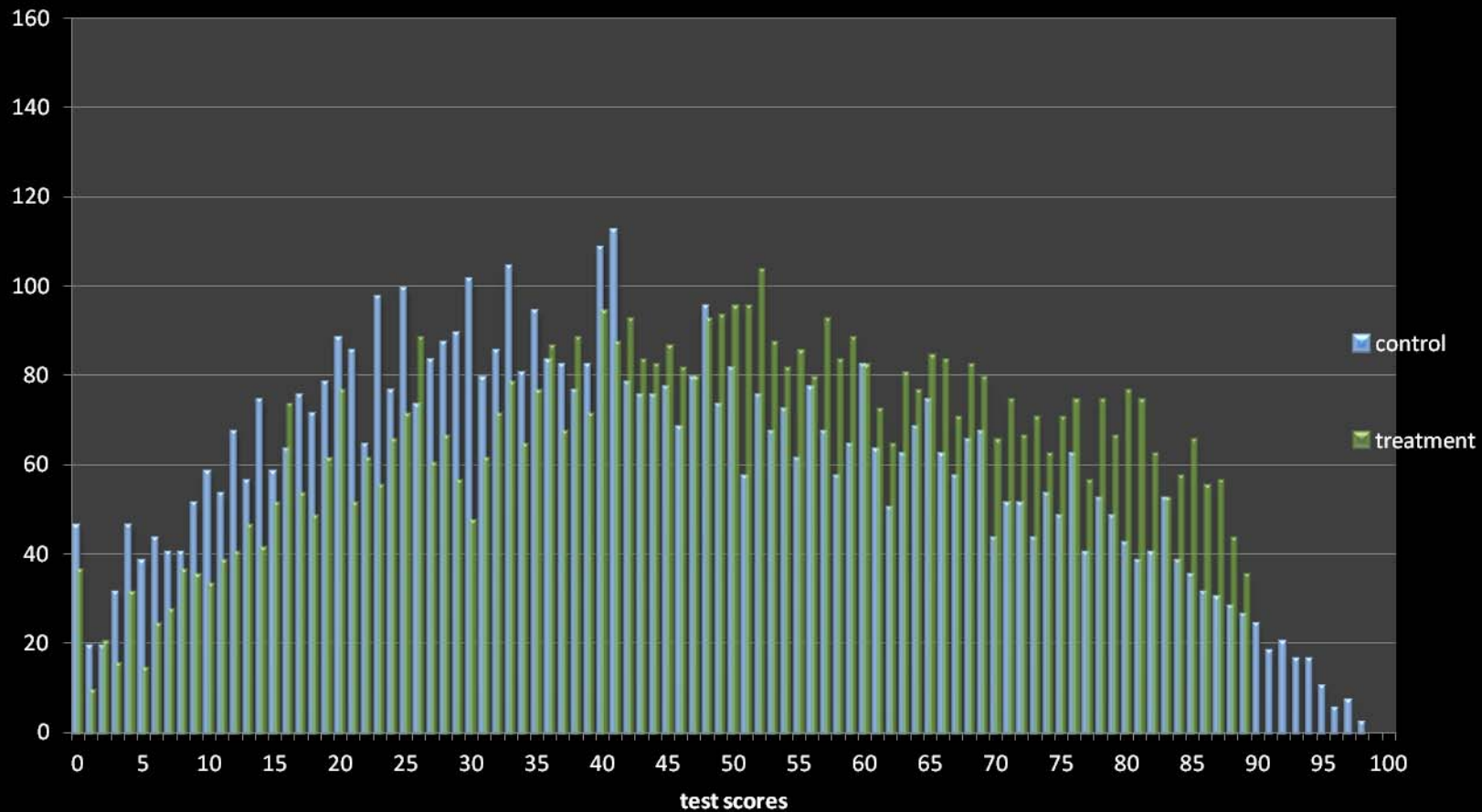
Endline test scores



Now, look at the improvement. Very few scored zero, and many scored much closer to the 40-point range...

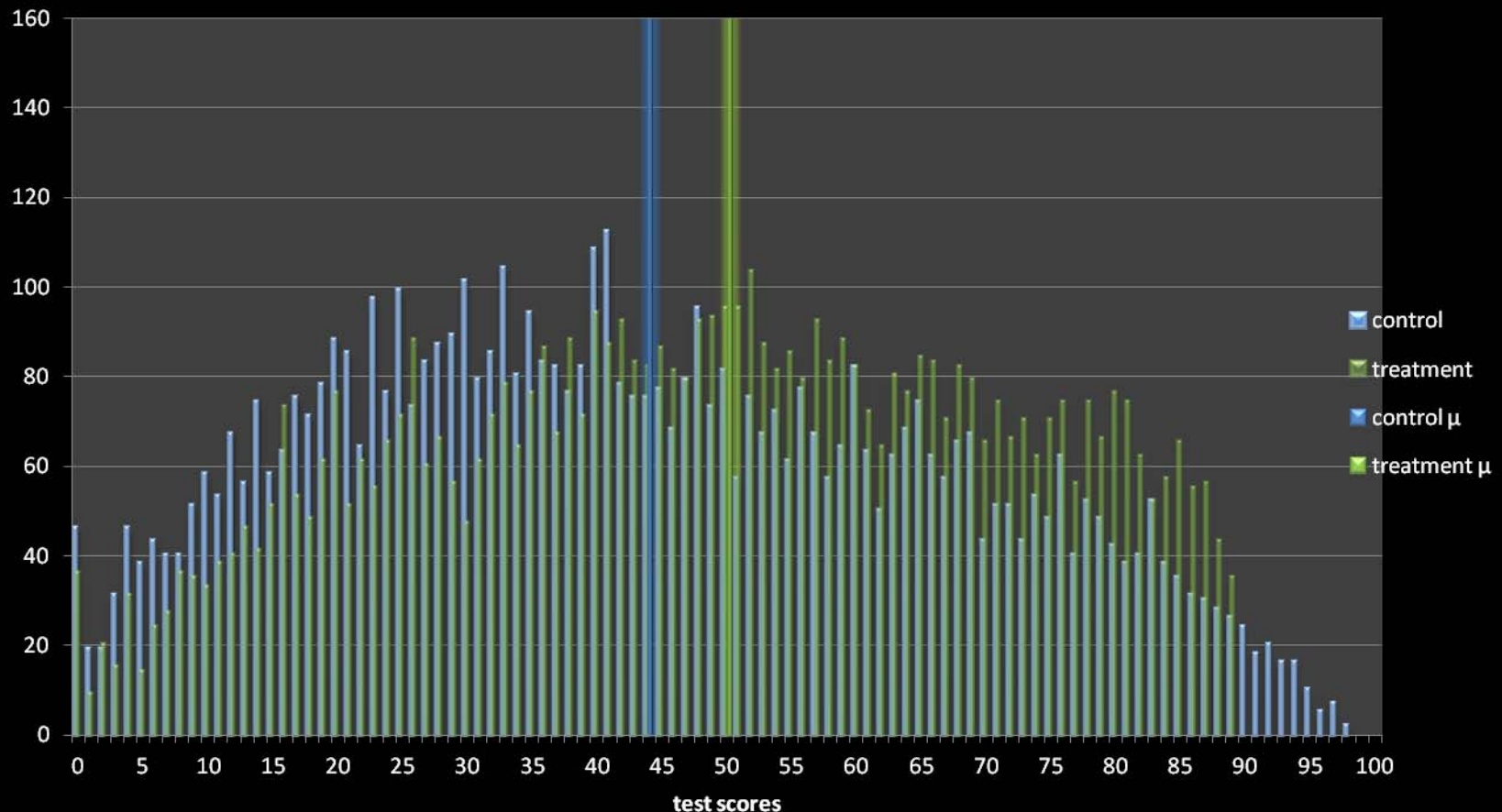
Was there an impact?

Post-test: control & treatment



Stop! That was the control group. The treatment group is green.

Average difference: 6 points



This is the true difference between the 2 groups

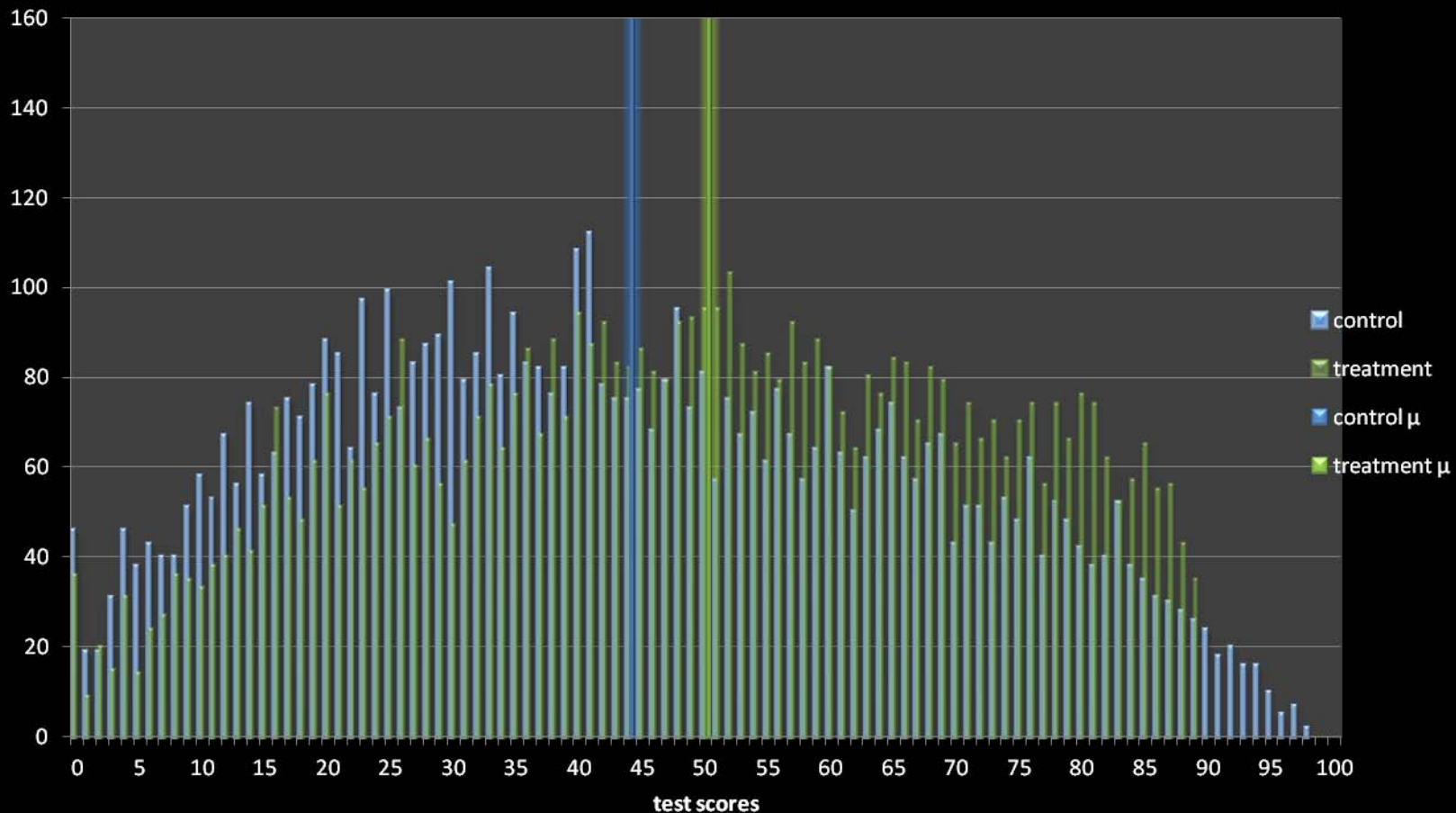
Population versus Sample

- Population: what we want to learn about
- Sample: what we see
 - How many children would we need to randomly sample to detect that the difference between the two groups is statistically significantly *different from zero*?

OR

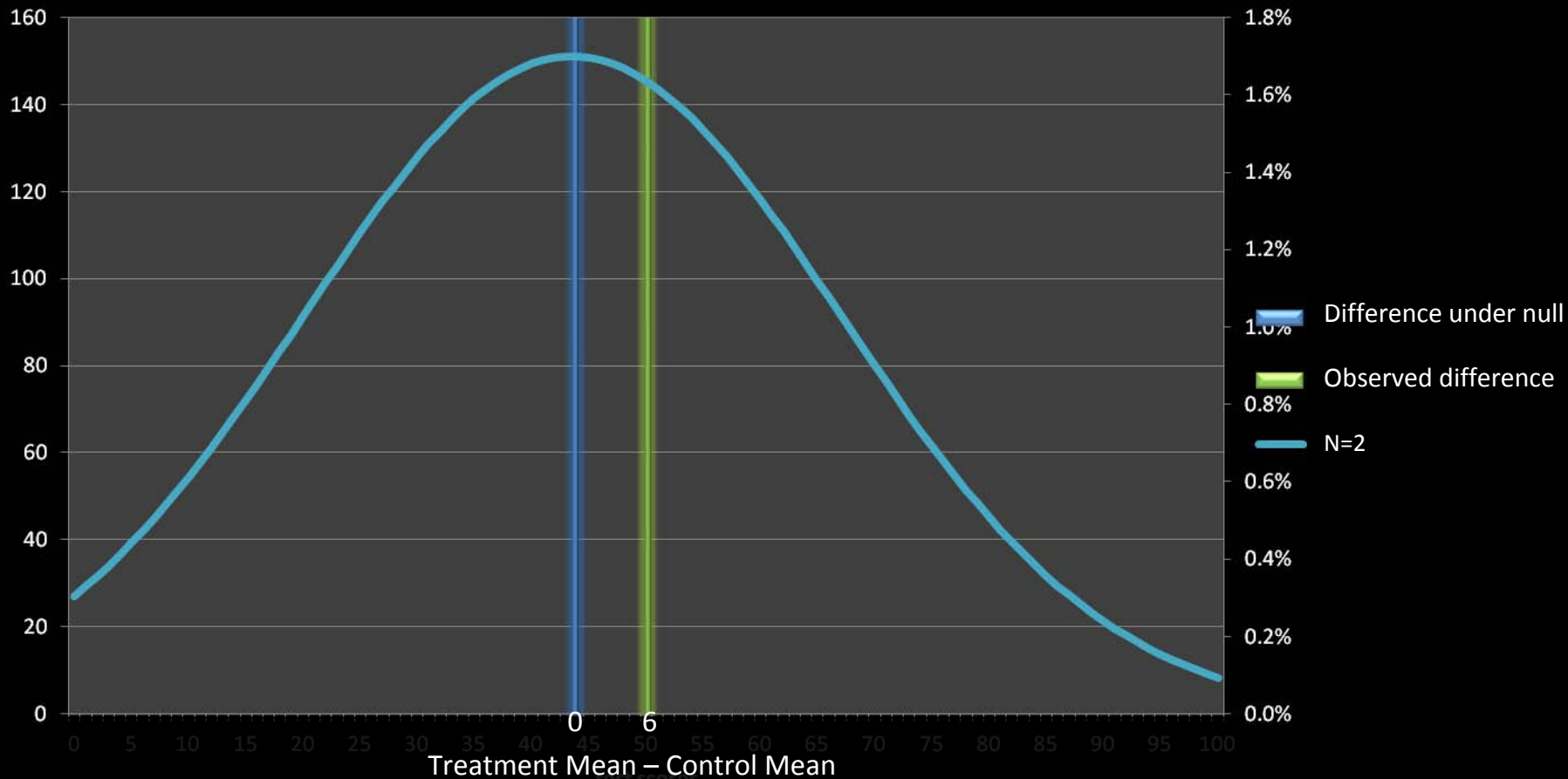
- How many children would we need to randomly sample to approximate *the true difference* with sufficient precision?

Testing statistical significance

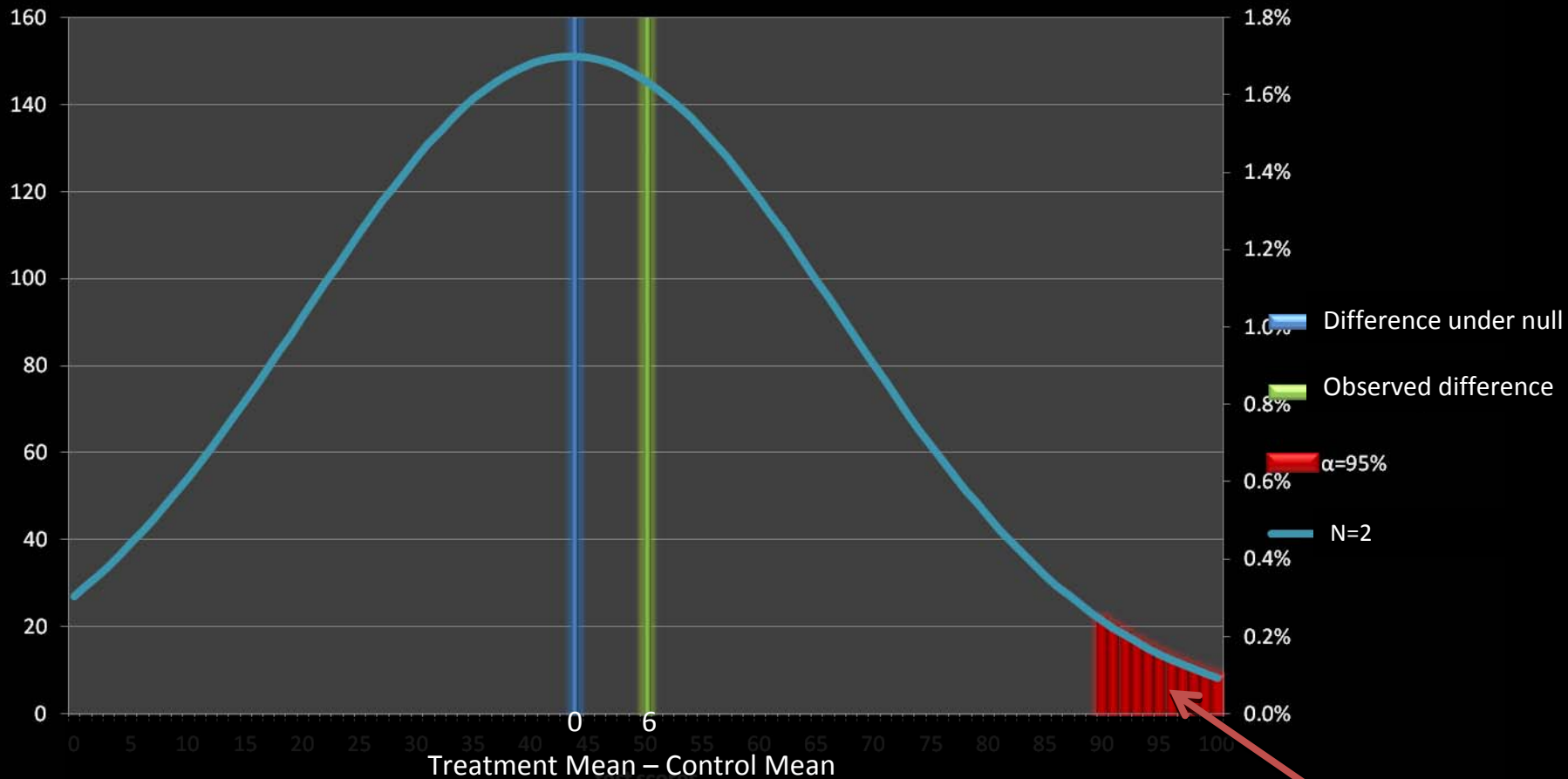


What's the probability that the 6 point difference is due to chance?

That probability depends on sample size (here: $N=2$)

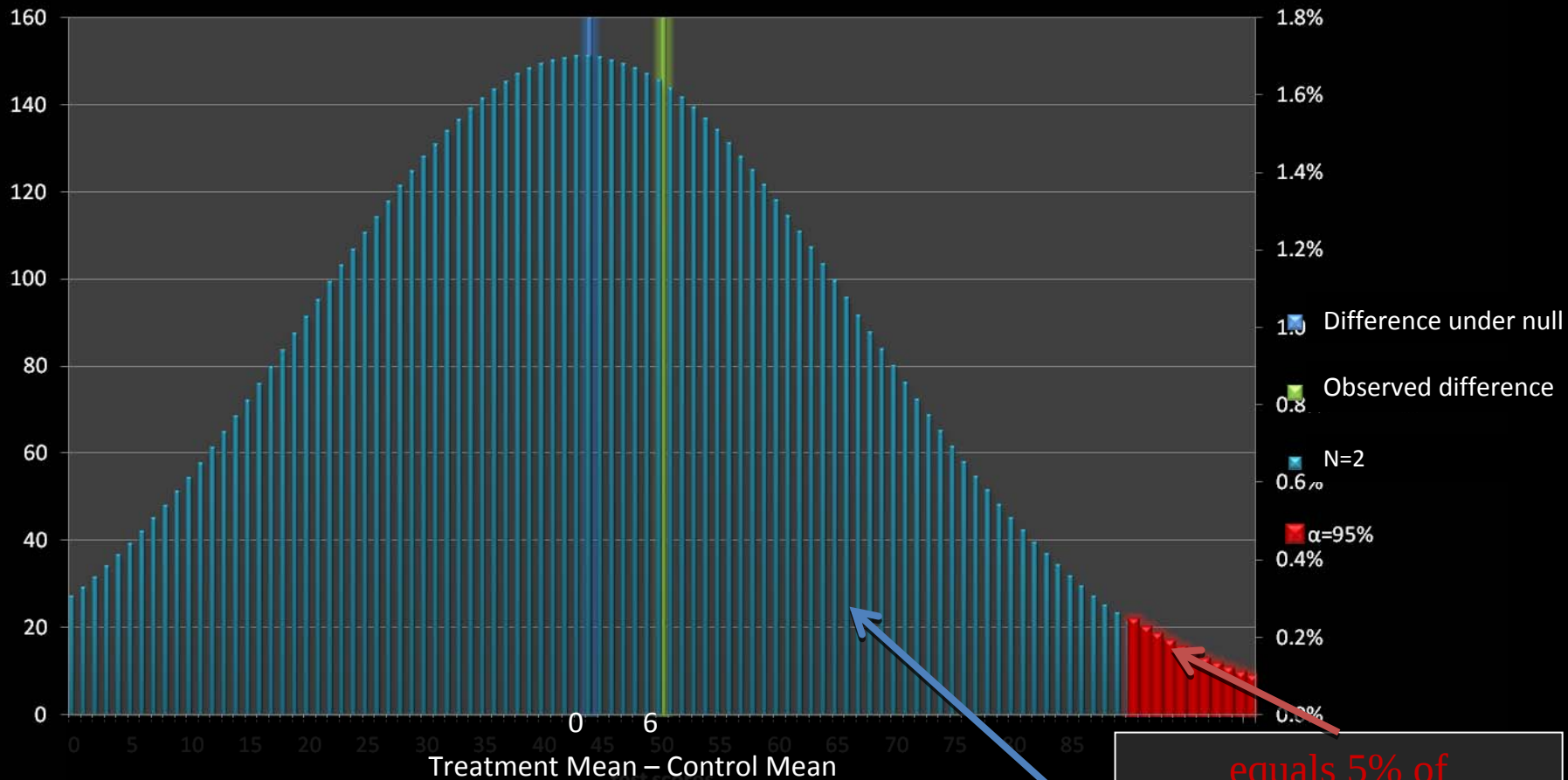


“Significance level” (5%)



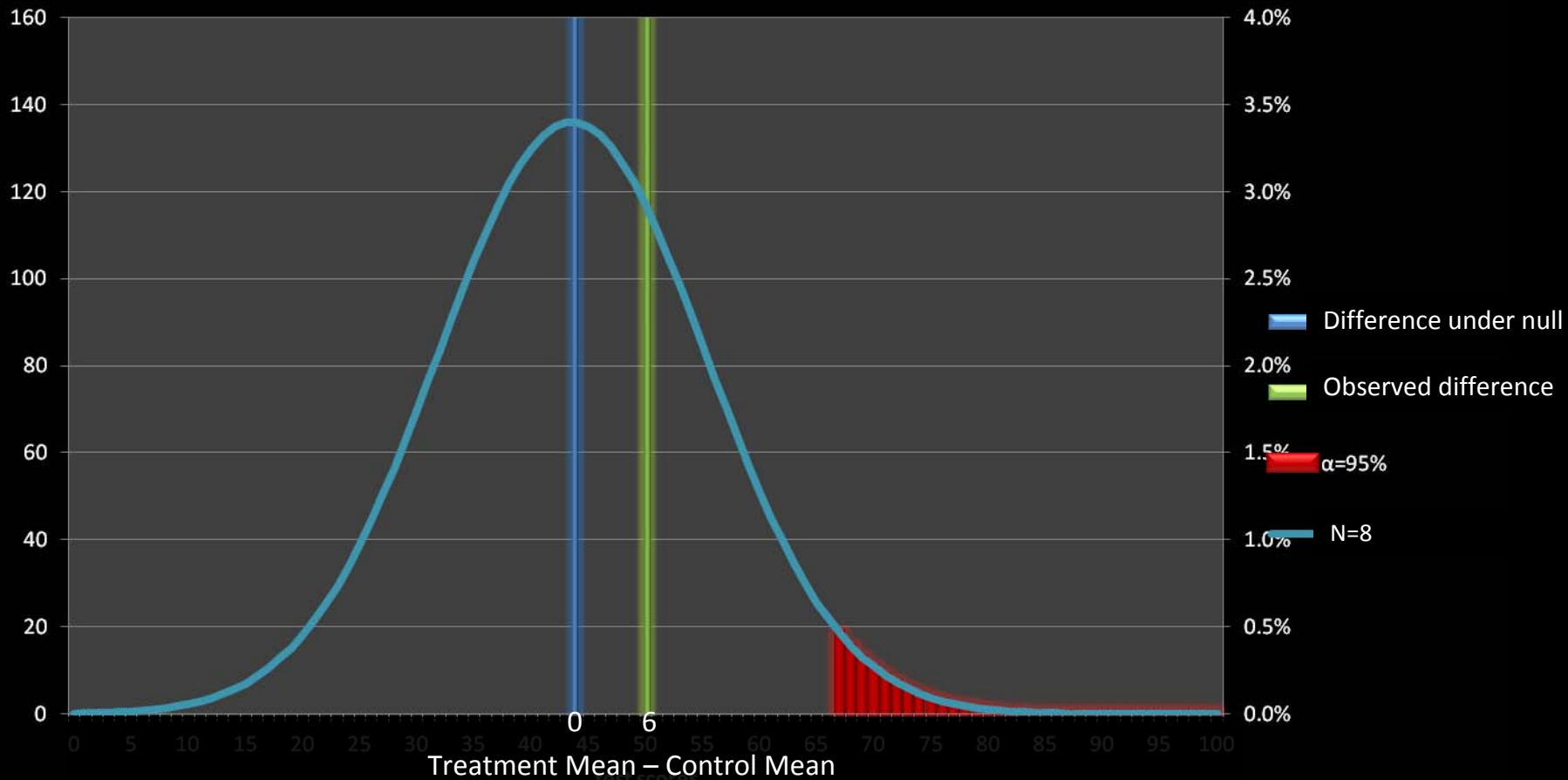
Critical region

“Significance level” (5%)

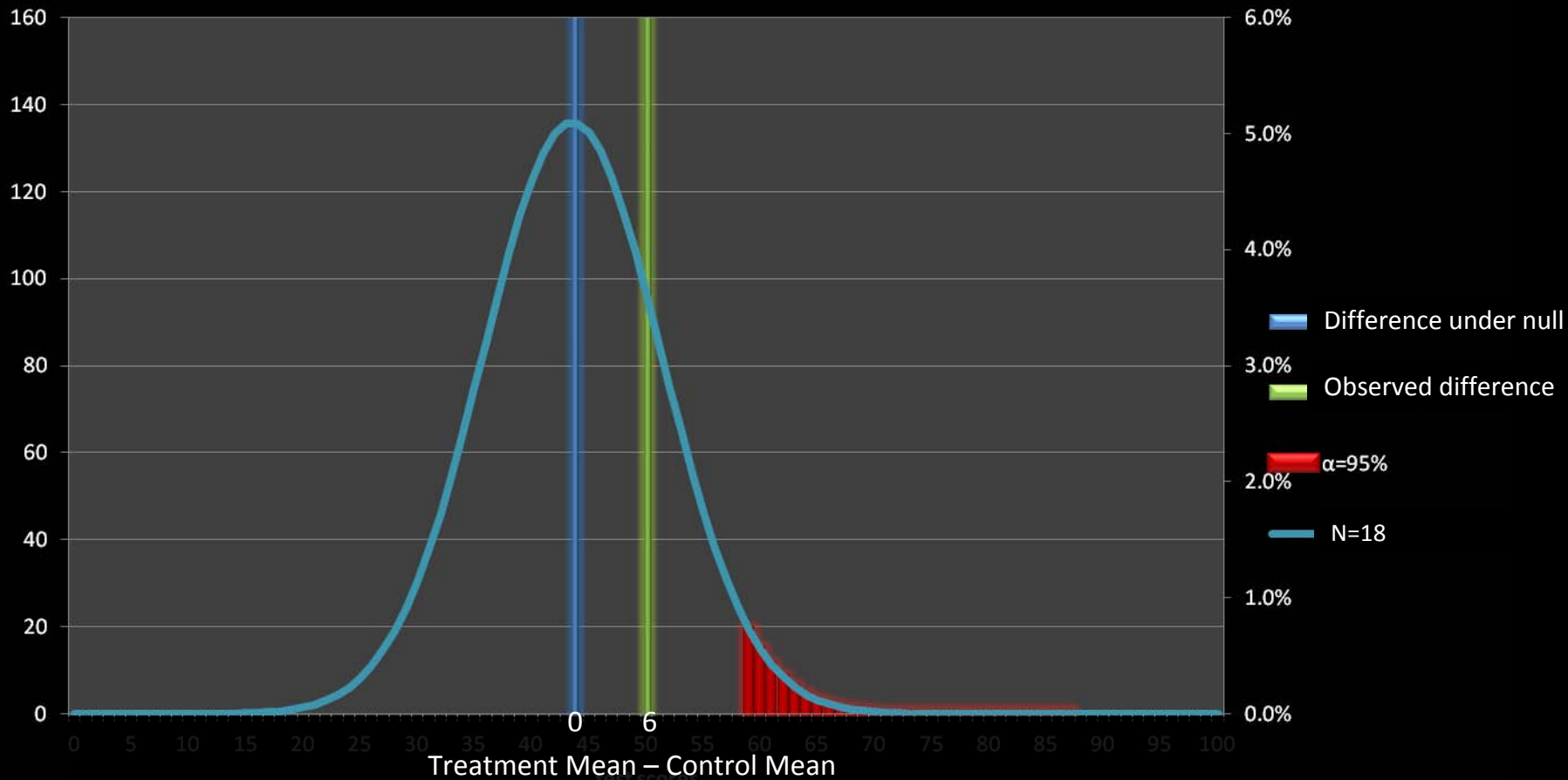


this total area equals 5% of

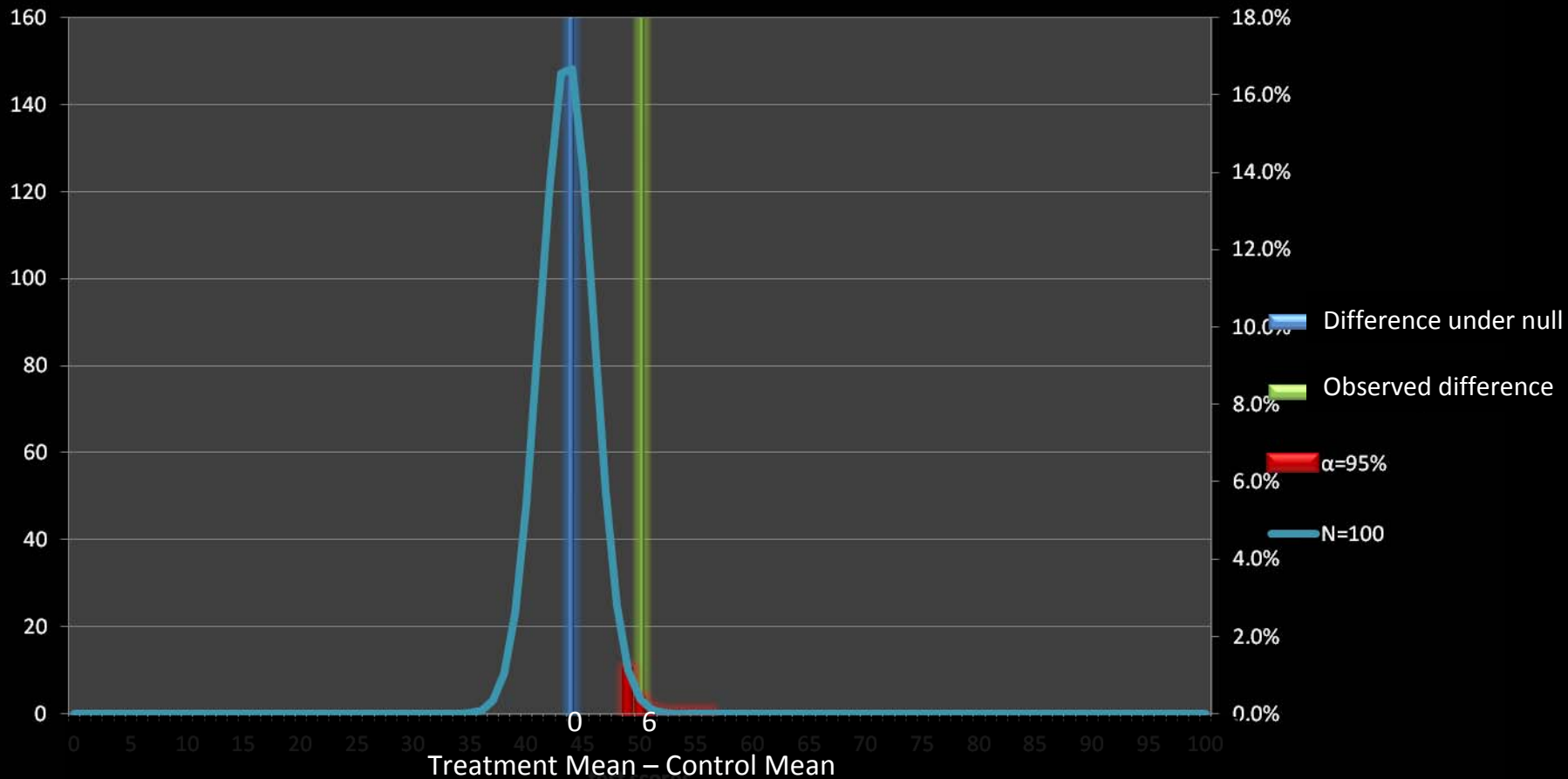
Significance: Sample size = 8



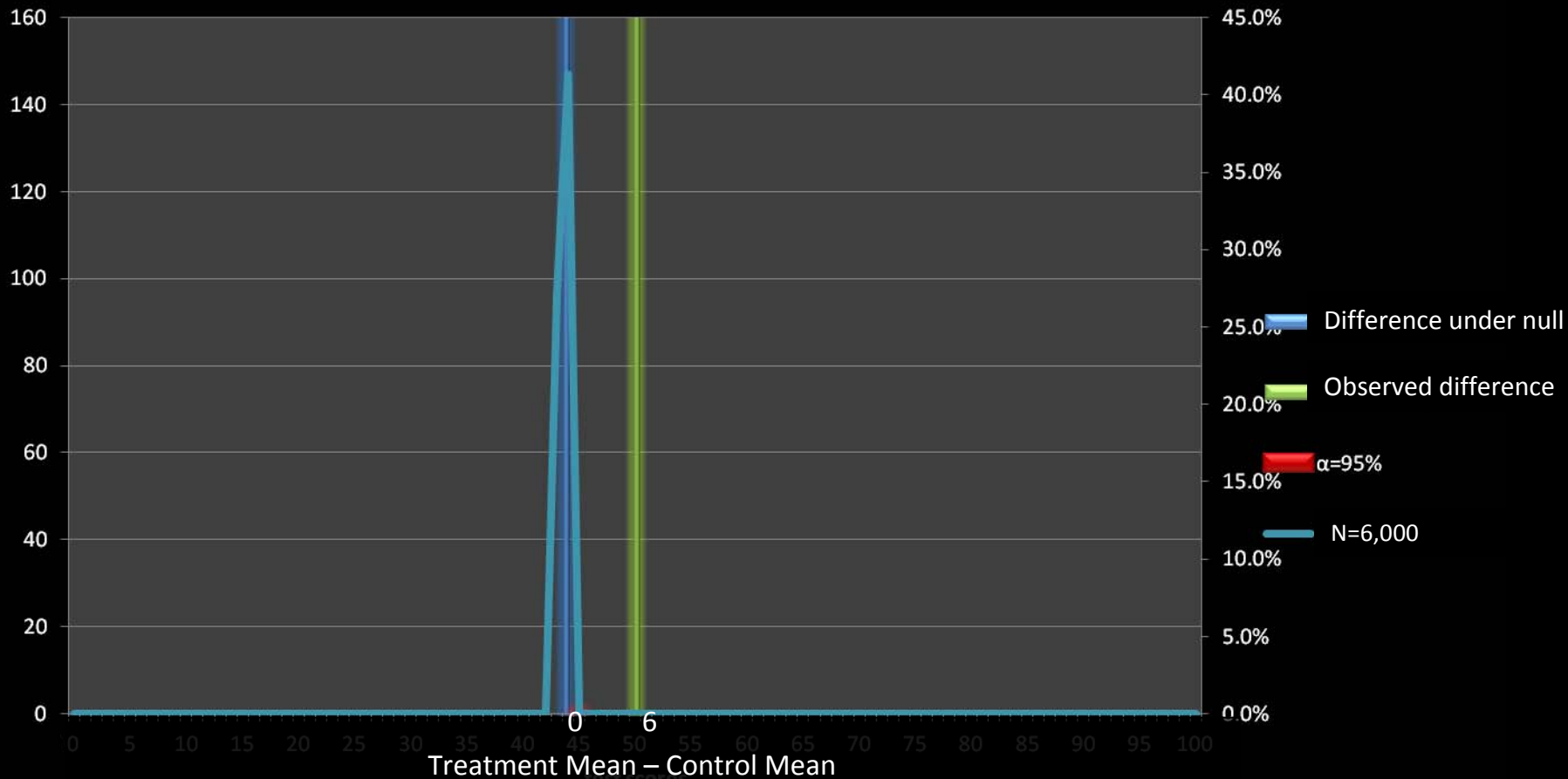
Significance: Sample size = 18



Significance: Sample size = 100



Significance: Sample size = 6,000



Hypothesis testing: conclusions

- What if the probability is greater than 5%?
 - We can't reject our null hypothesis
 - Are we 100 percent certain there is no impact?
 - No, it just didn't meet the statistical threshold to conclude otherwise
 - Perhaps there is indeed no impact
 - Or perhaps there is impact,
 - But not enough sample to detect it most of the time
 - Or we got a very unlucky sample this time
 - How do we reduce this error?

POWER!

Hypothesis testing: conclusions

- When we use a “95% confidence interval”
- How frequently will we “detect” effective programs?
- That is Statistical Power

Hypothesis testing: 95% confidence

		YOU CONCLUDE	
		<i>Effective</i>	<i>No Effect</i>
THE TRUTH	<i>Effective</i>		Type II Error (low power) 
	<i>No Effect</i>	Type I Error (5% of the time) 	

Power:

- How frequently will we “detect” effective programs?

Power: main ingredients

1. Variance

- The more “noisy” it is to start with, the harder it is to measure effects

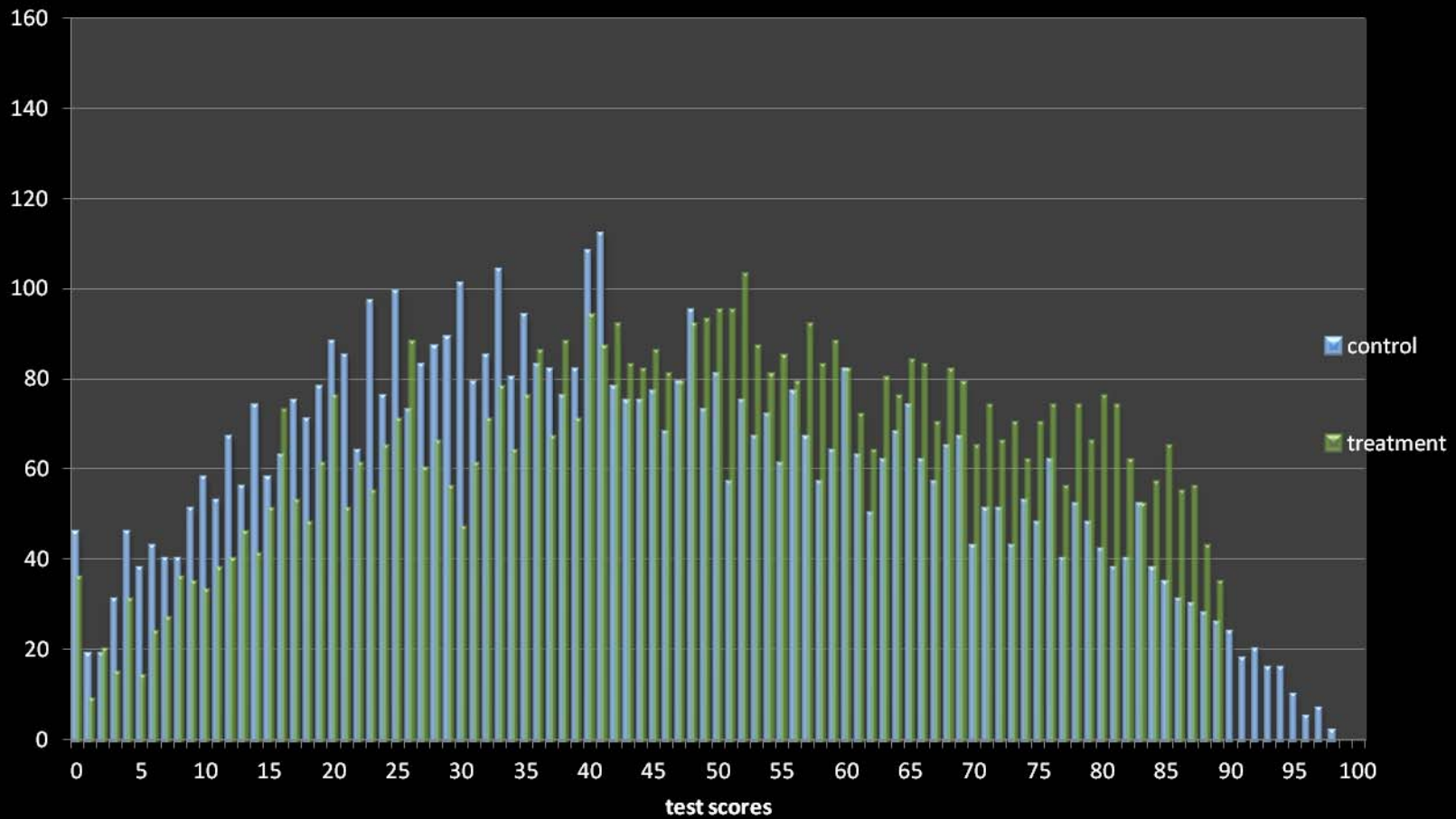
2. Effect Size to be detected

- The more fine (or more precise) the effect size we want to detect, the larger sample we need
- Smallest effect size with practical / policy significance?

3. Sample Size

- The more children we sample, the more likely we are to obtain the true difference

Variance



Variance

- There is very little we can do to reduce the noise
- The underlying variance is what it is
- We can try to “absorb” variance:
 - using a baseline
 - controlling for other variables

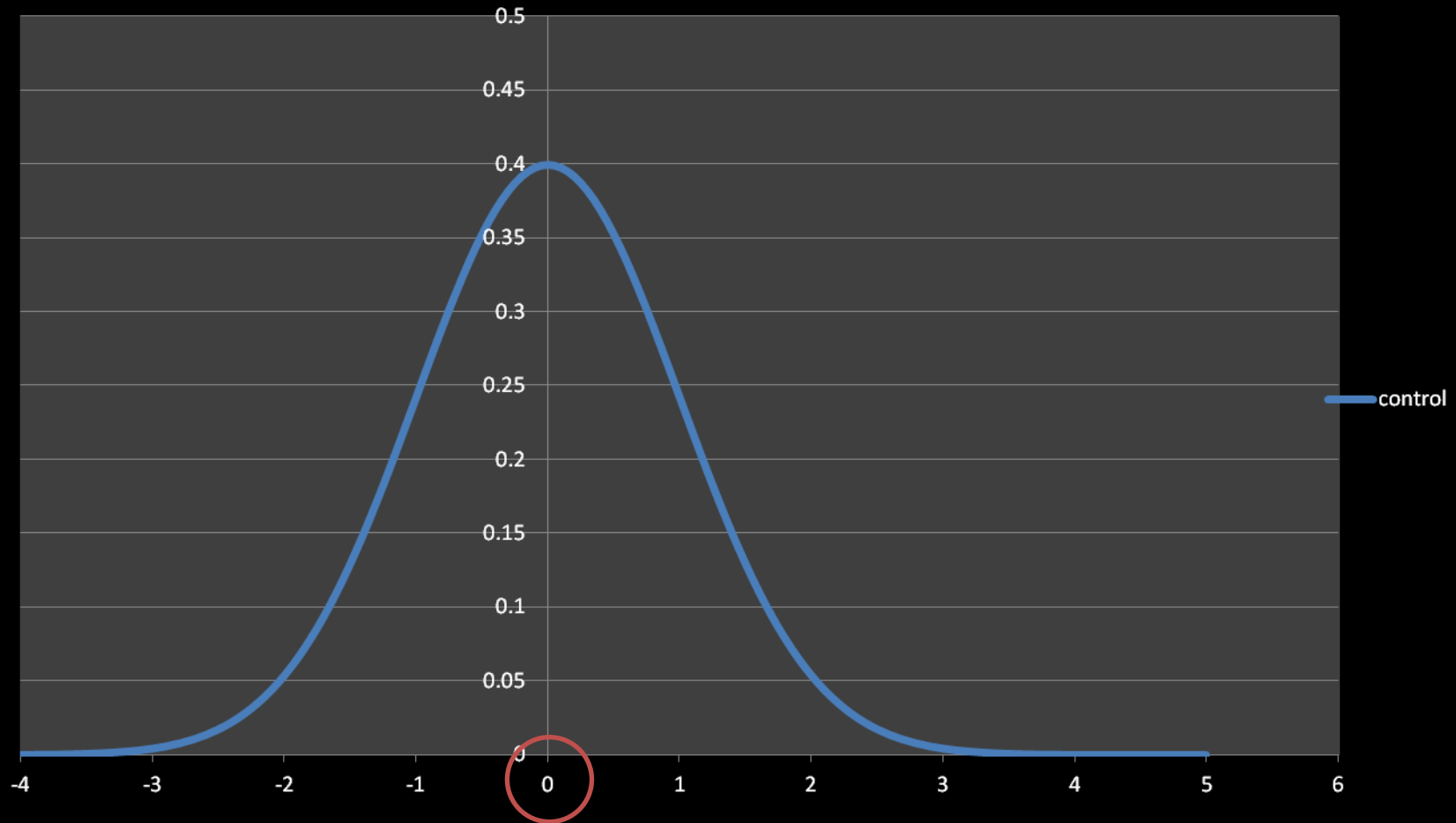
Effect Size

- To calculate statistical significance we start with the “null hypothesis”:
- To think about statistical power, we need to propose a secondary hypothesis

2 Hypotheses & “significance level”

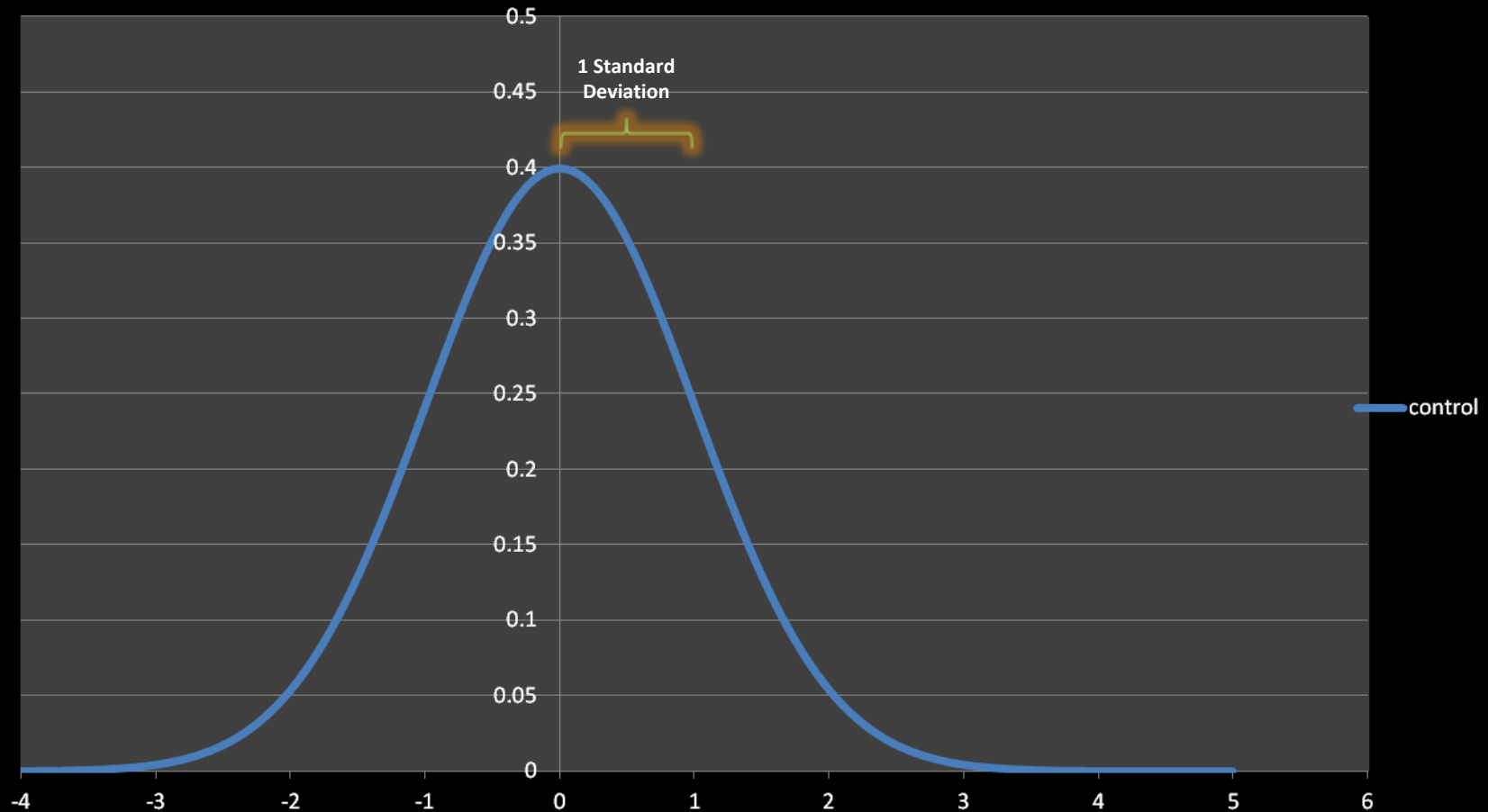
- The following is an example...

Null Hypothesis: assume zero impact



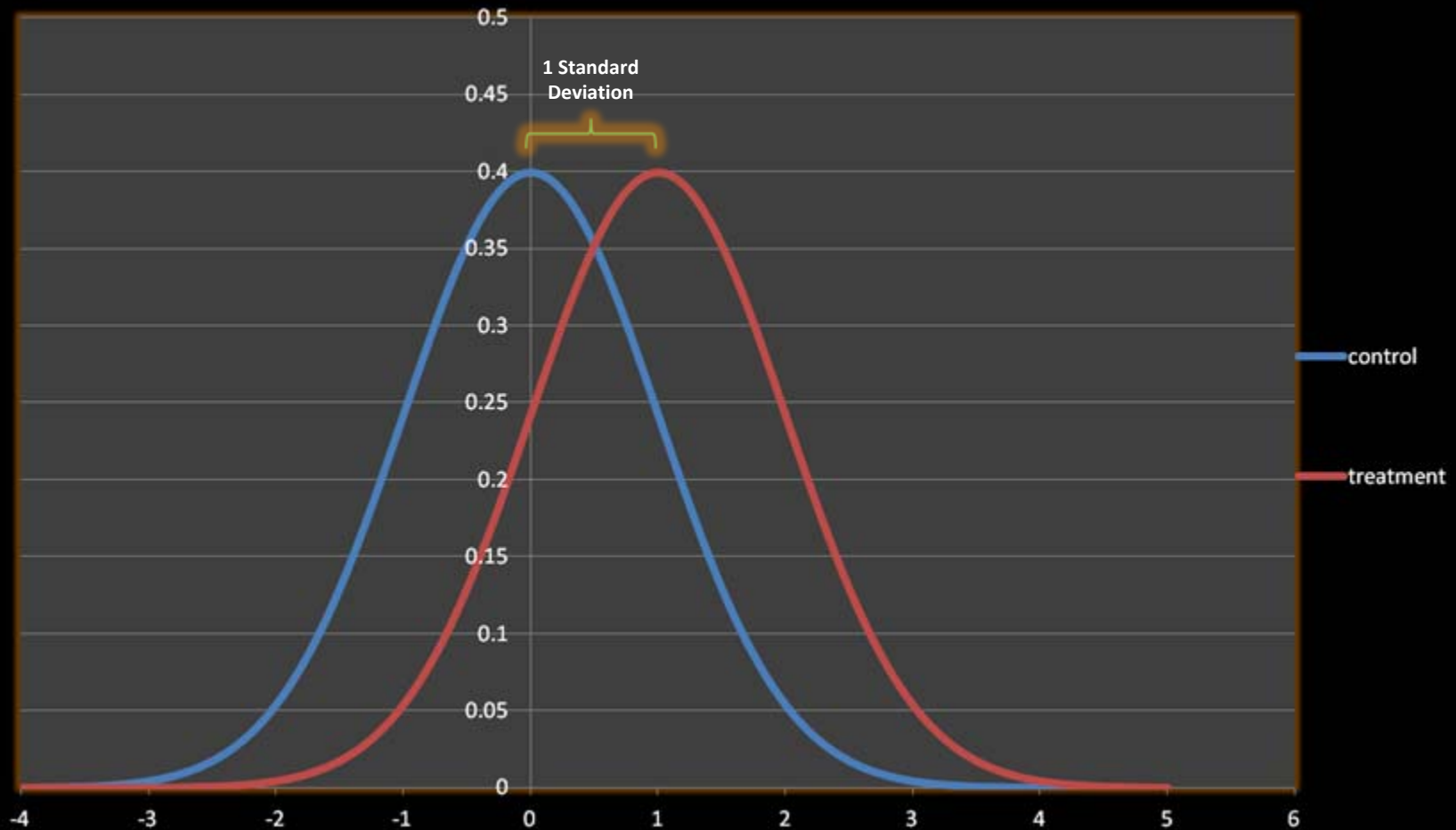
“Impact = 0” There’s a sampling distribution around that.

Effect Size: 1 “standard deviation”



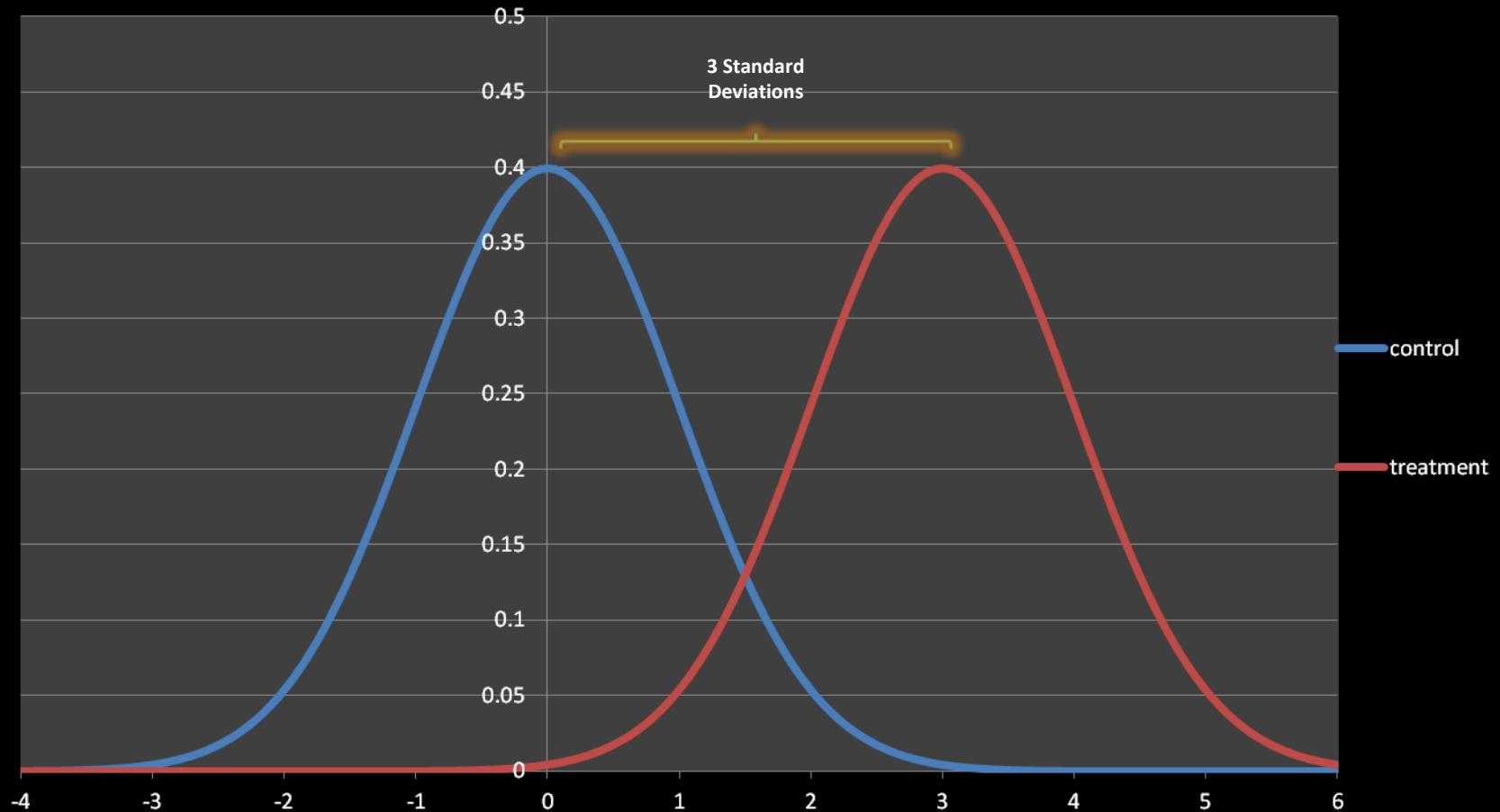
We hypothesize another possible “true effect size”

Effect Size: 1 “standard deviation”



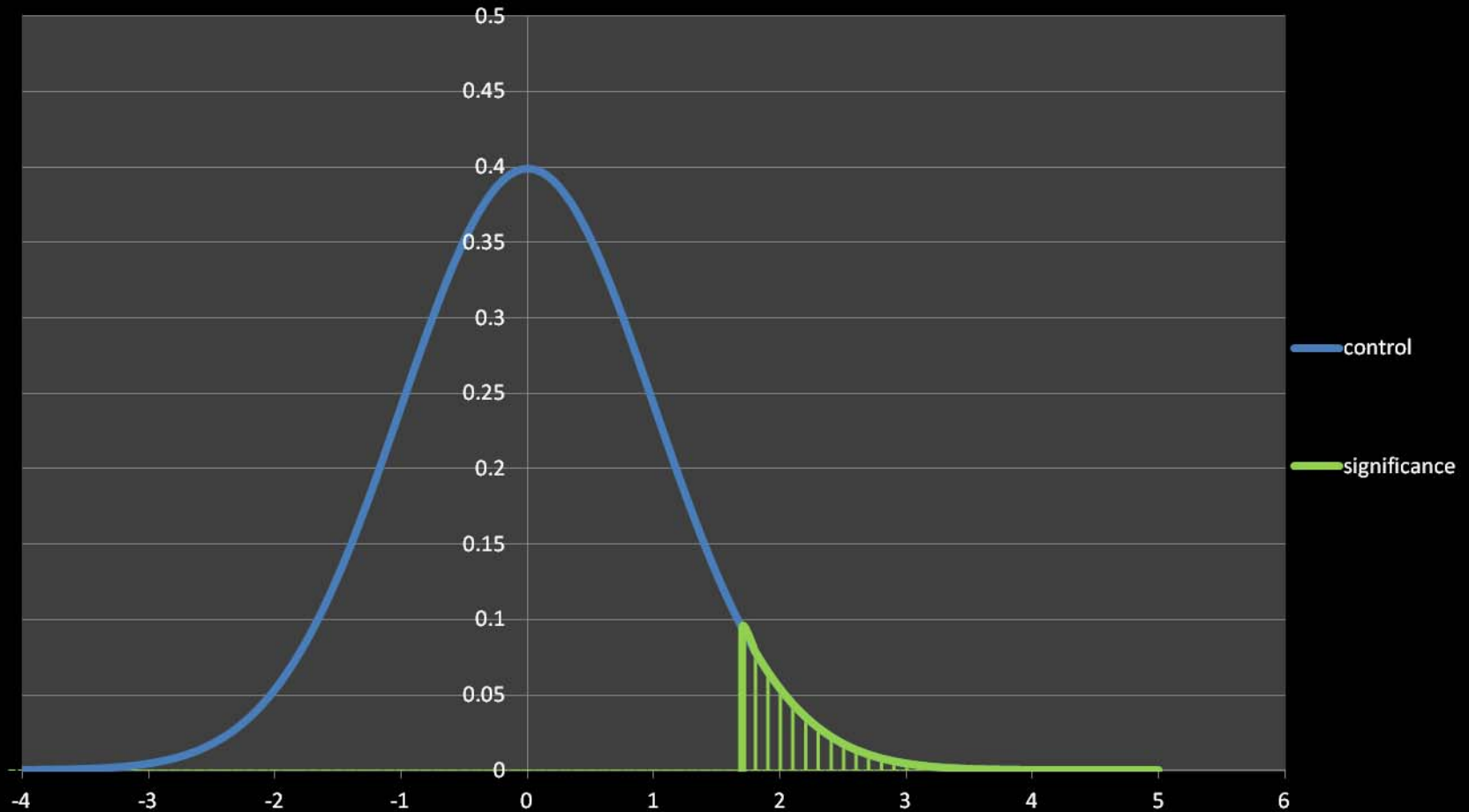
And there's a new sampling distribution around that

Effect Size: 3 standard deviations

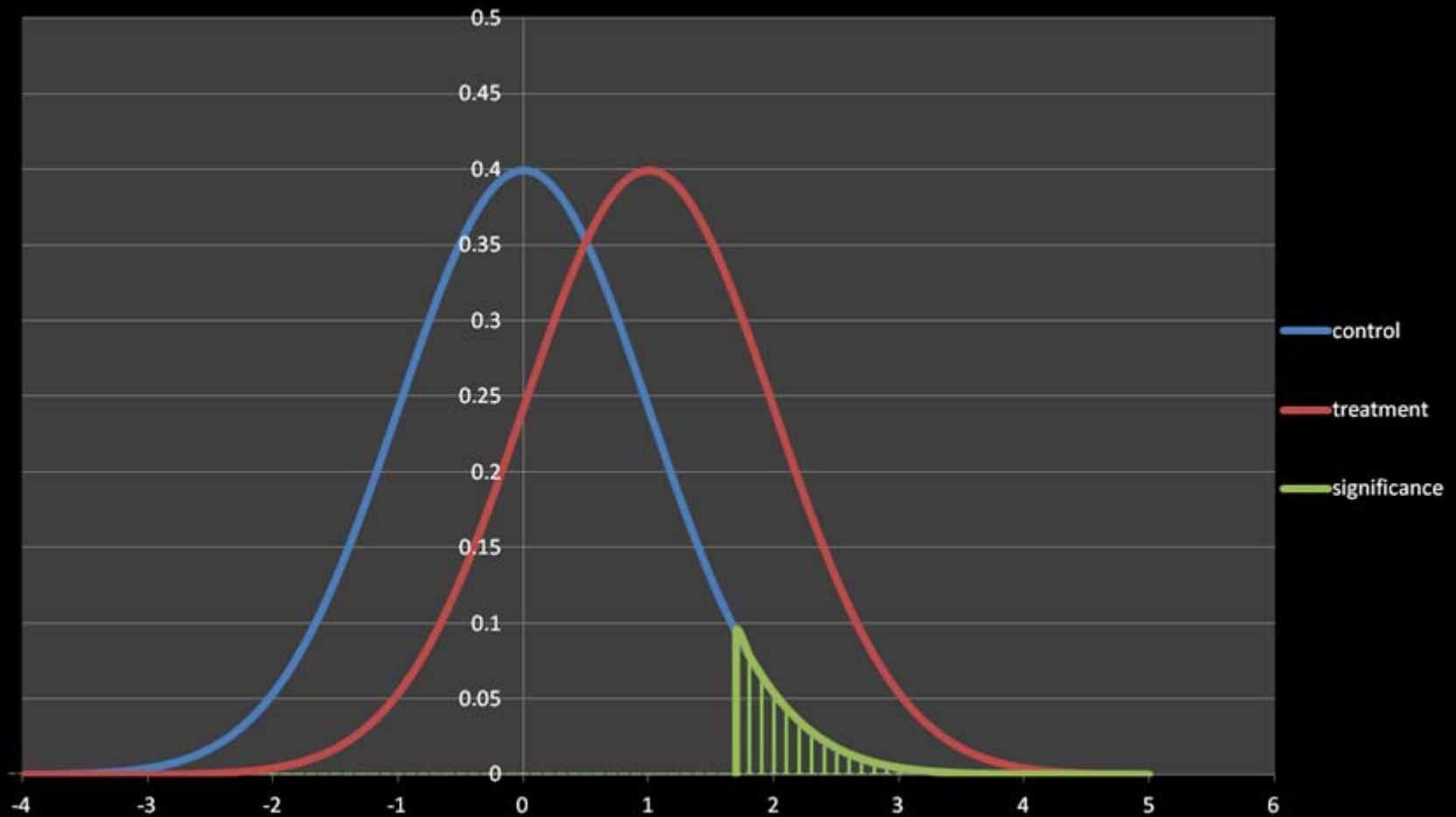


The less overlap the better...

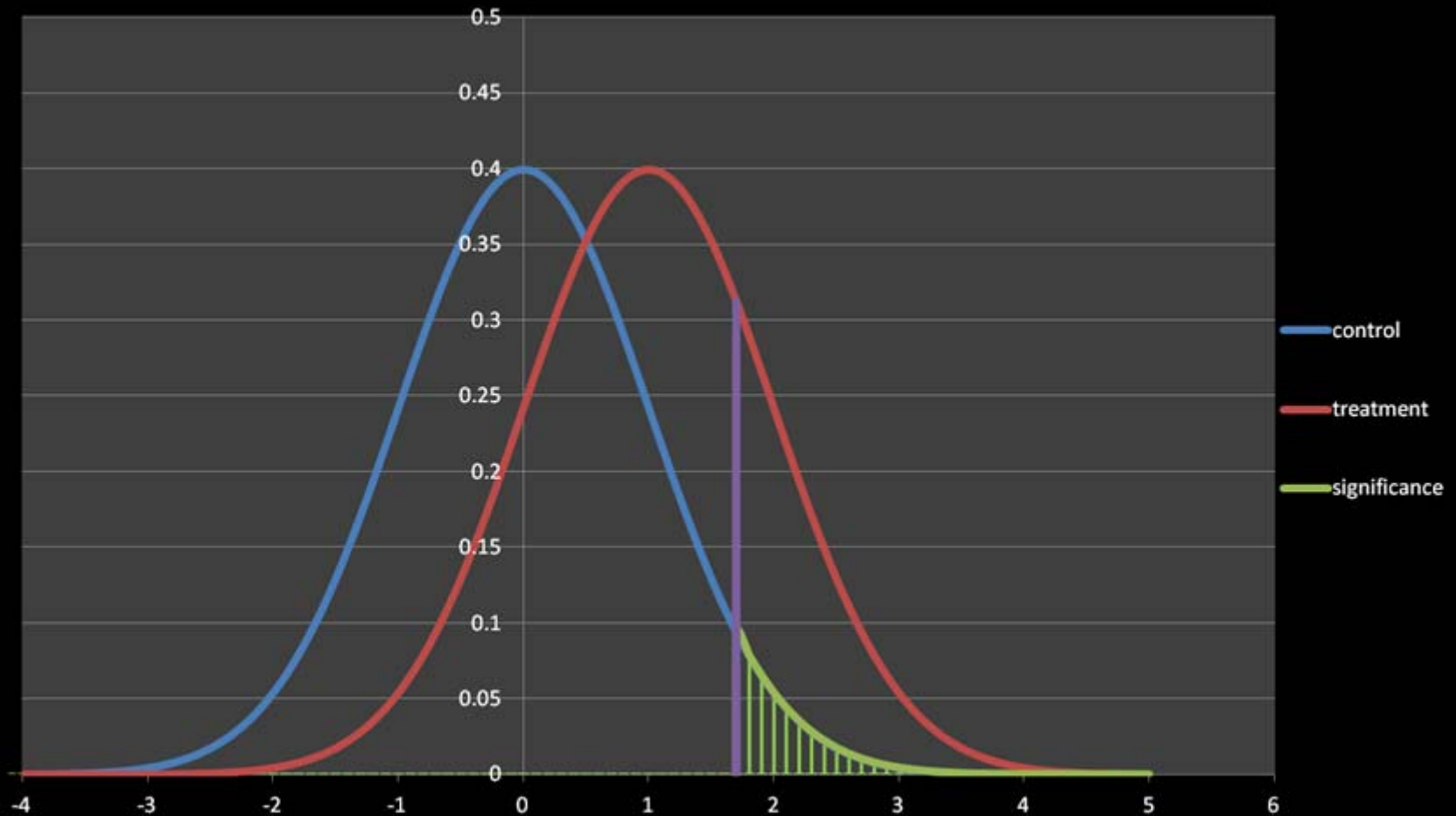
Significance level: reject H_0 in critical region



True effect is 1 SD

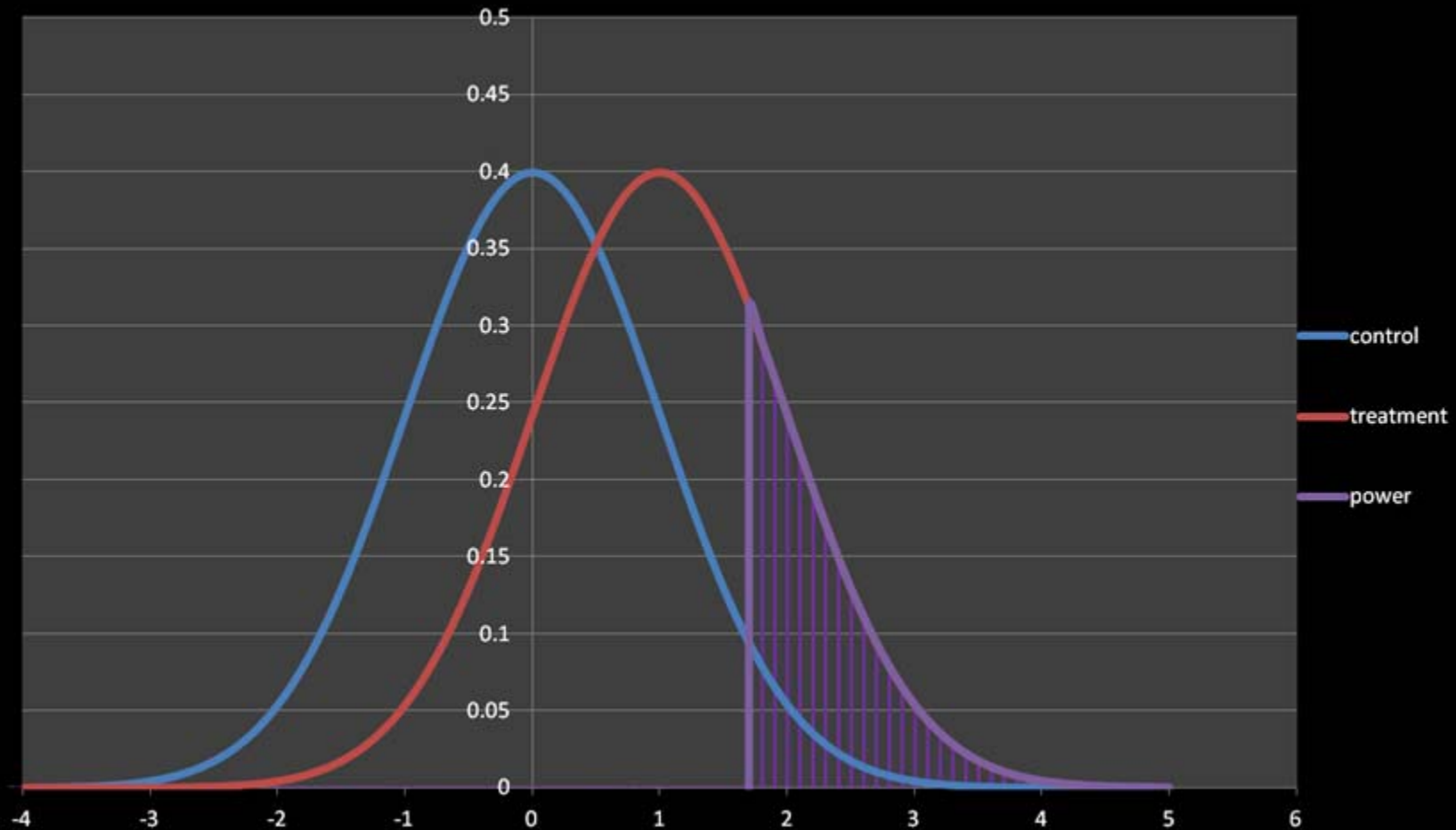


Power: when is H_0 rejected?



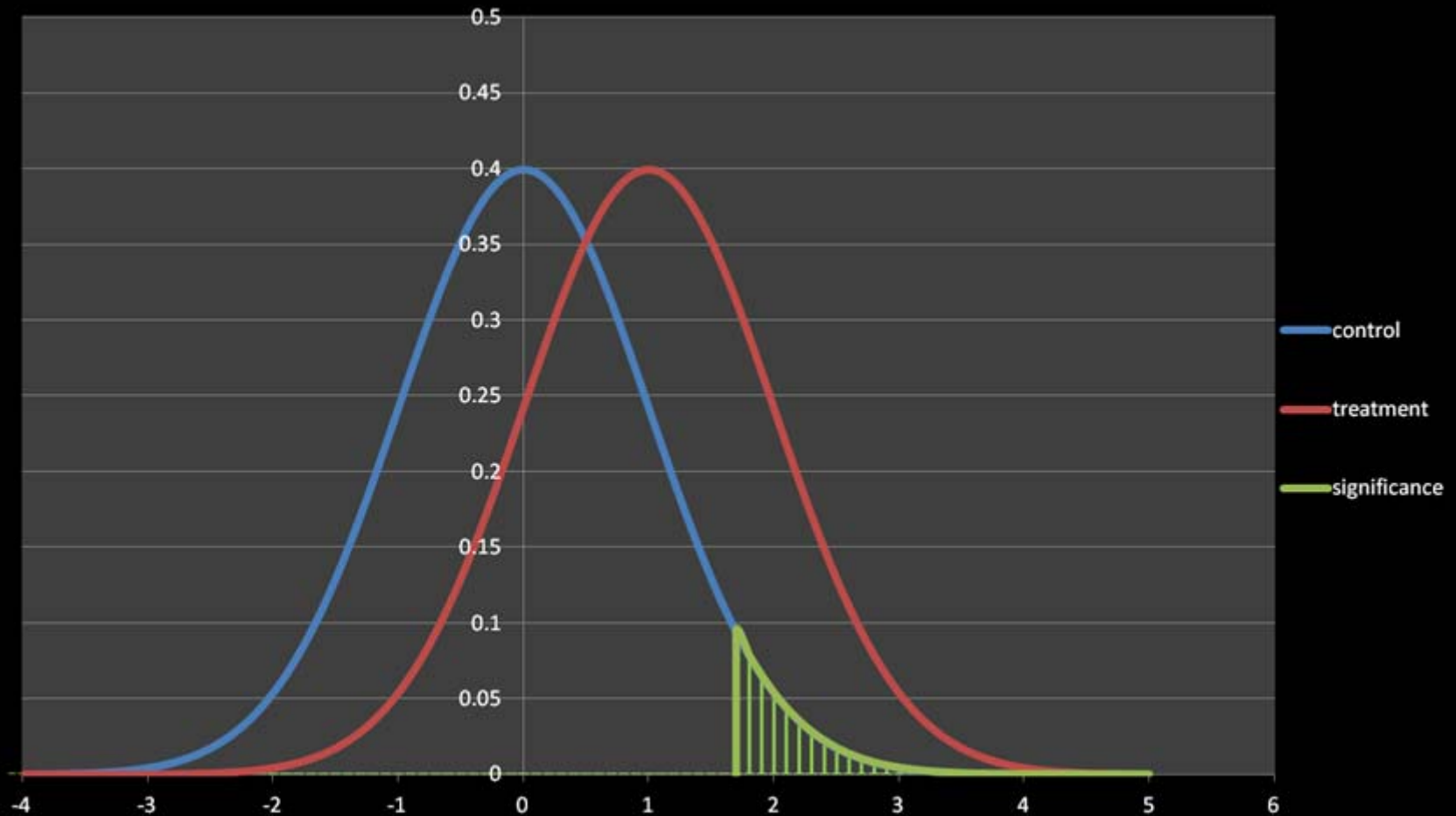
Power: 26%

If the true impact was 1SD...

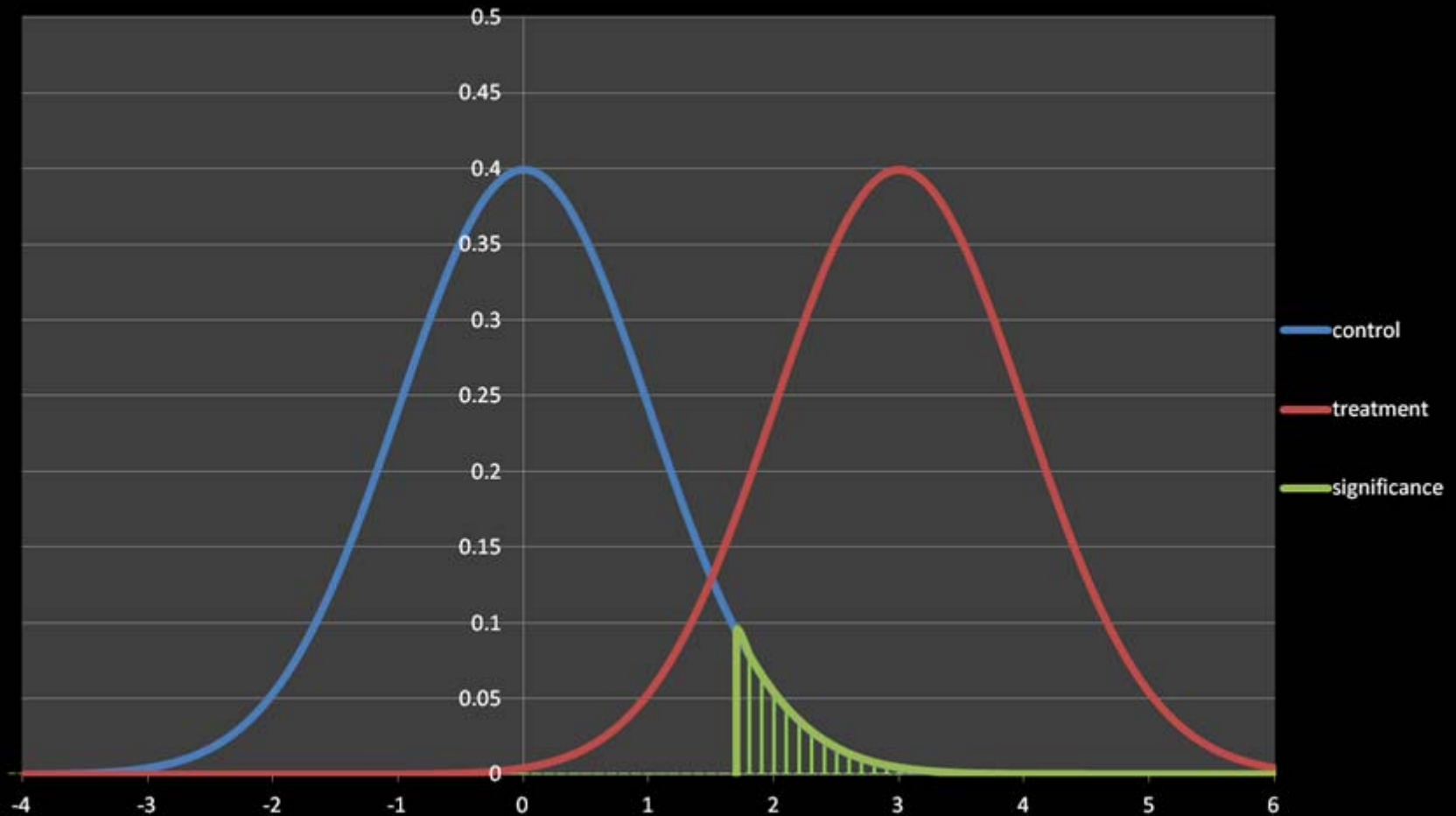


The Null Hypothesis would be rejected only 26% of the time

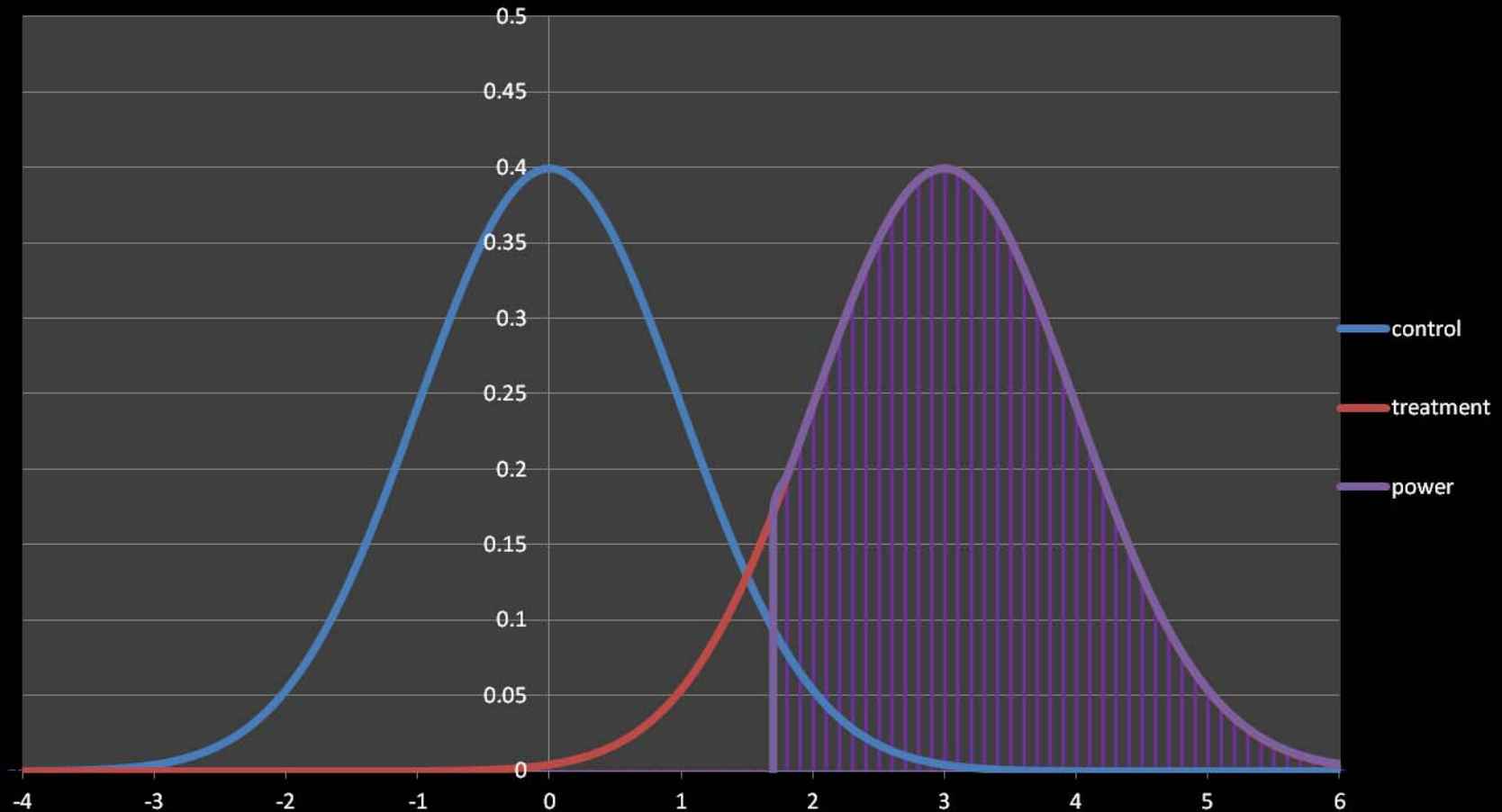
Power: if we change the effect size?



Power: assume effect size = 3 SDs



Power: 91%



The Null Hypothesis would be rejected 91% of the time

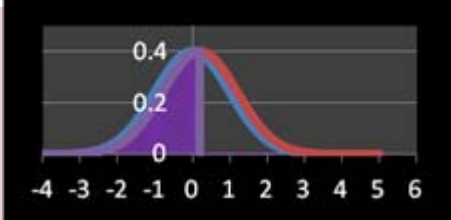
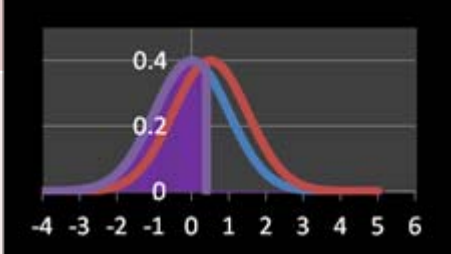
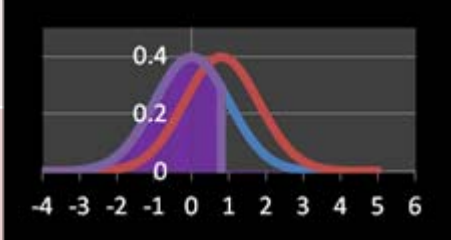
Picking an effect size

- What is the smallest effect that should justify the program being adopted?
- If the effect is smaller than that, it might as well be zero: we are not interested in proving that a very small effect is different from zero
- In contrast, if any effect larger than that would justify adopting this program: we want to be able to distinguish it from zero

Standardized effect sizes

- How large an effect you can detect with a given sample depends on how variable the outcome is.
- The Standardized effect size is the effect size divided by the standard deviation of the outcome
- Common effect sizes

Standardized effect size

An effect size of...	Is considered...	...and it means that...	
0.2	Modest	The average member of the treatment group had a better outcome than the 58 th percentile of the control group	
0.5	Large	The average member of the treatment group had a better outcome than the 69 th percentile of the control group	
0.8	VERY Large	The average member of the treatment group had a better outcome than the 79 th percentile of the control group	

Effect Size: Bottom Line

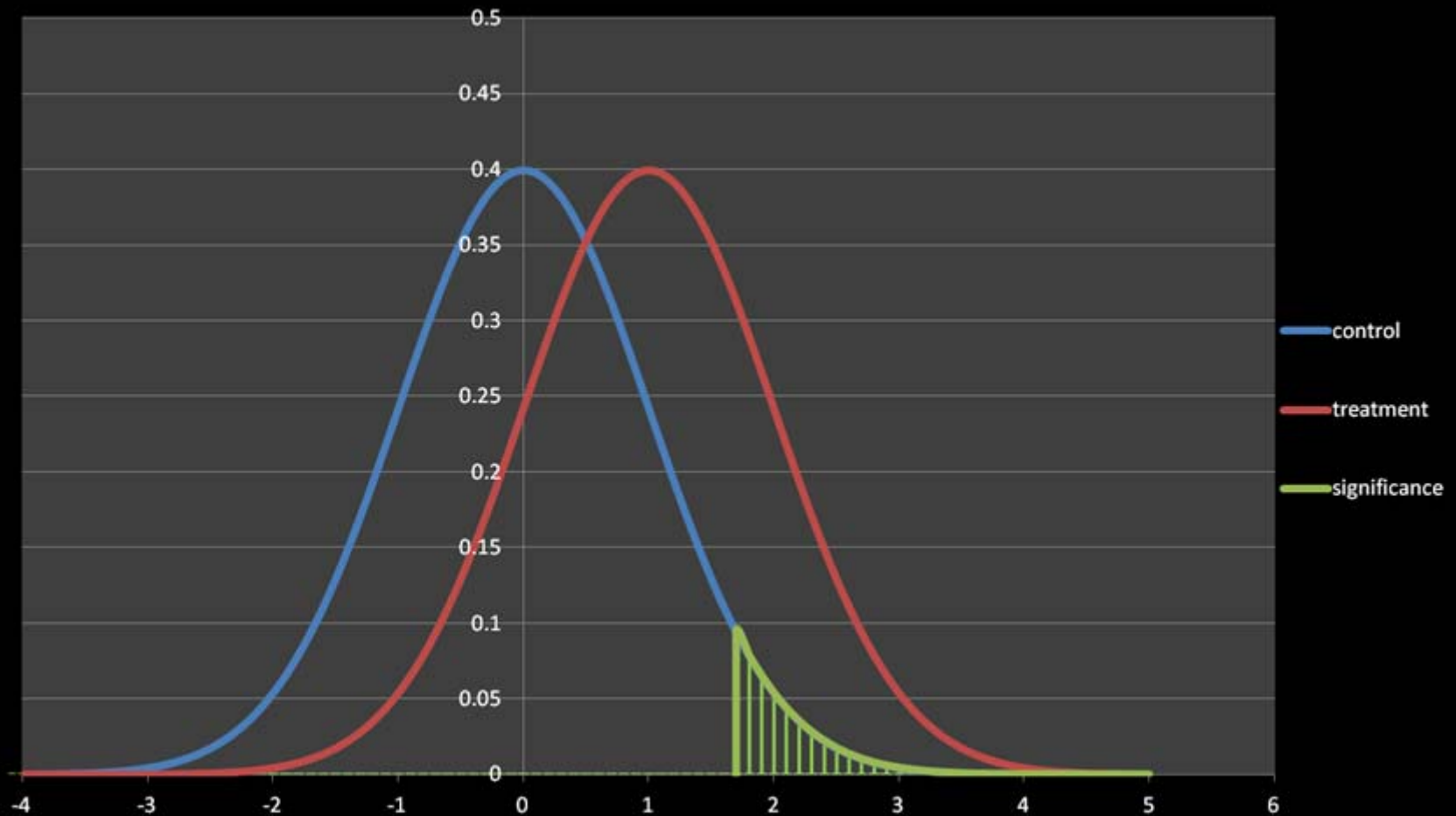
- You should not alter the effect size to achieve power
- The effect size is more of a policy question
- One variable that can affect effect size is take-up!
 - If your job training program increases income by 20%
 - But only $\frac{1}{2}$ of the people in your treatment group participate
 - You need to adjust your impact estimate accordingly
 - From 20% to 10%
- So how do you increase power?

Try: Increasing the sample size

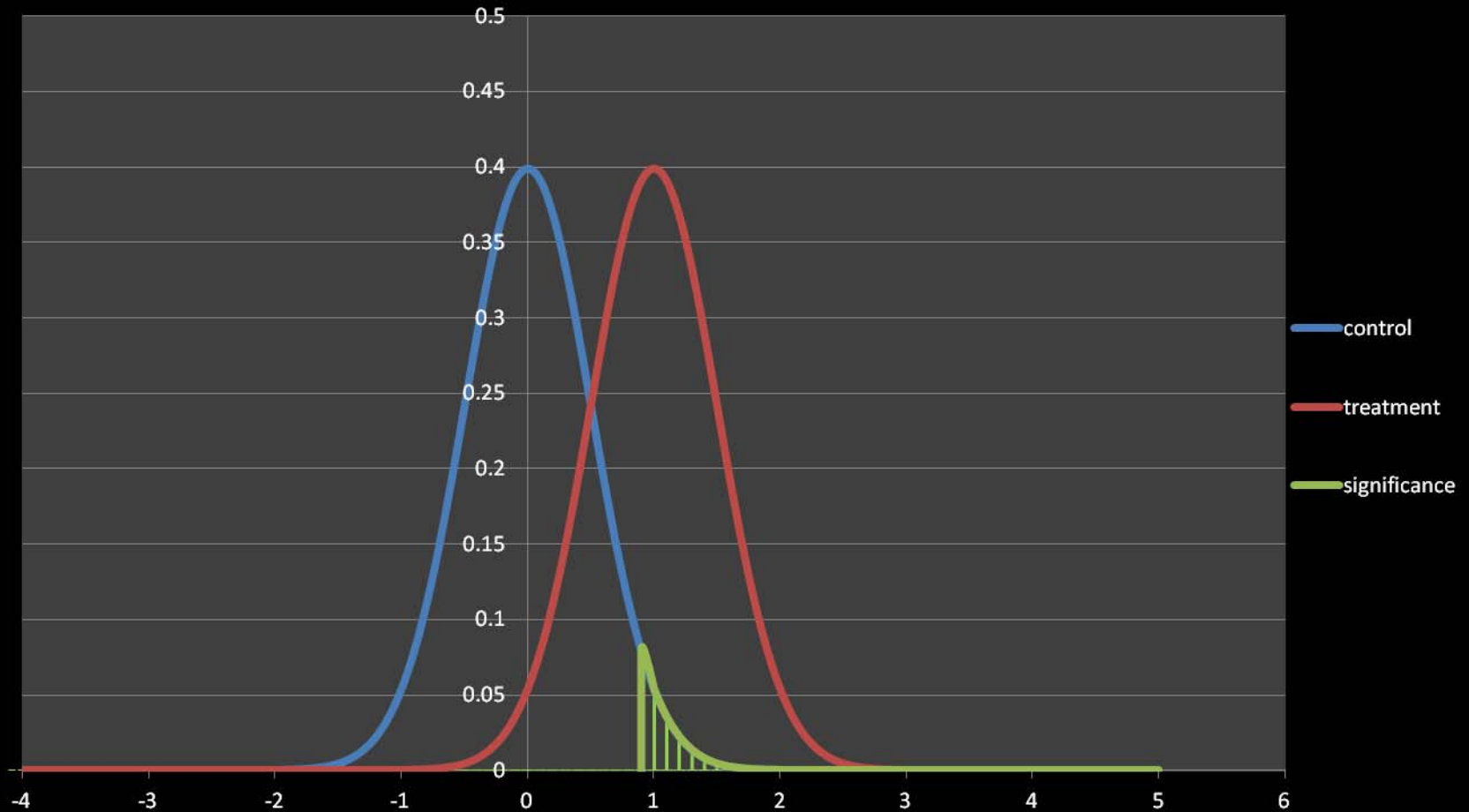
Sample size

- Increasing sample size reduces the “spread” of our bell curve
- The more observations we randomly pull, the more likely we get the “true average”

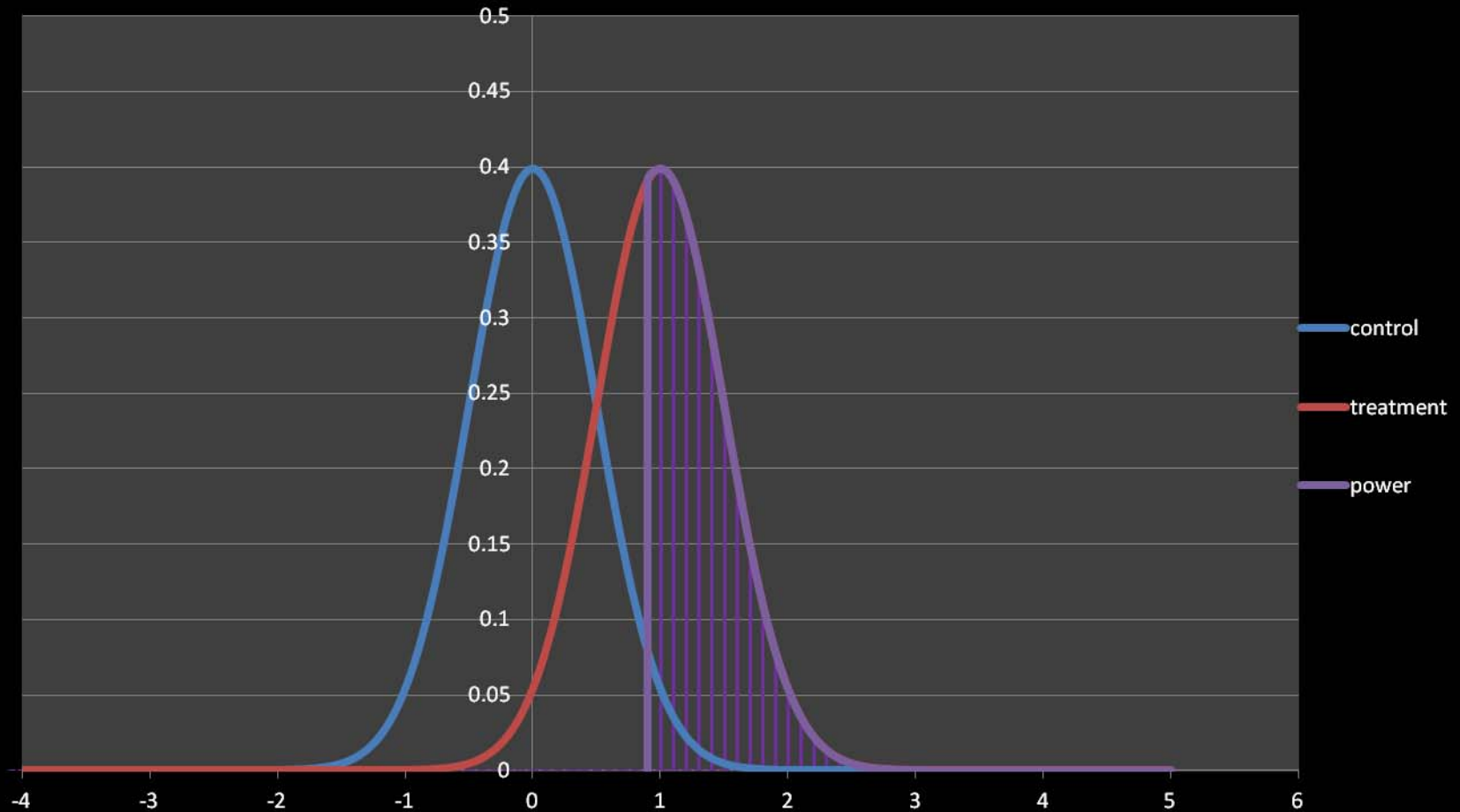
Power: Effect size = 1SD, Sample size = 1



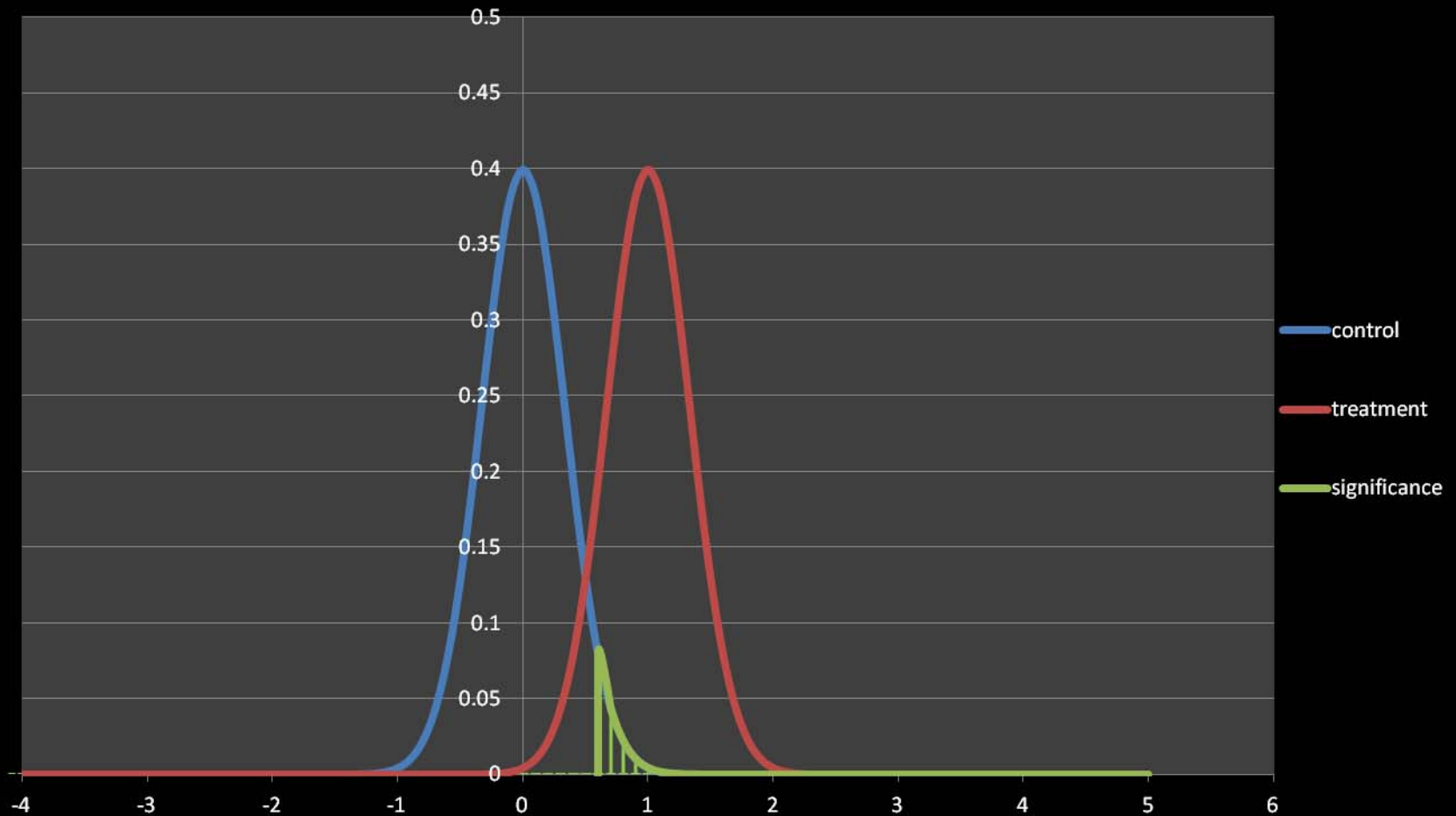
Power: Sample size = 4



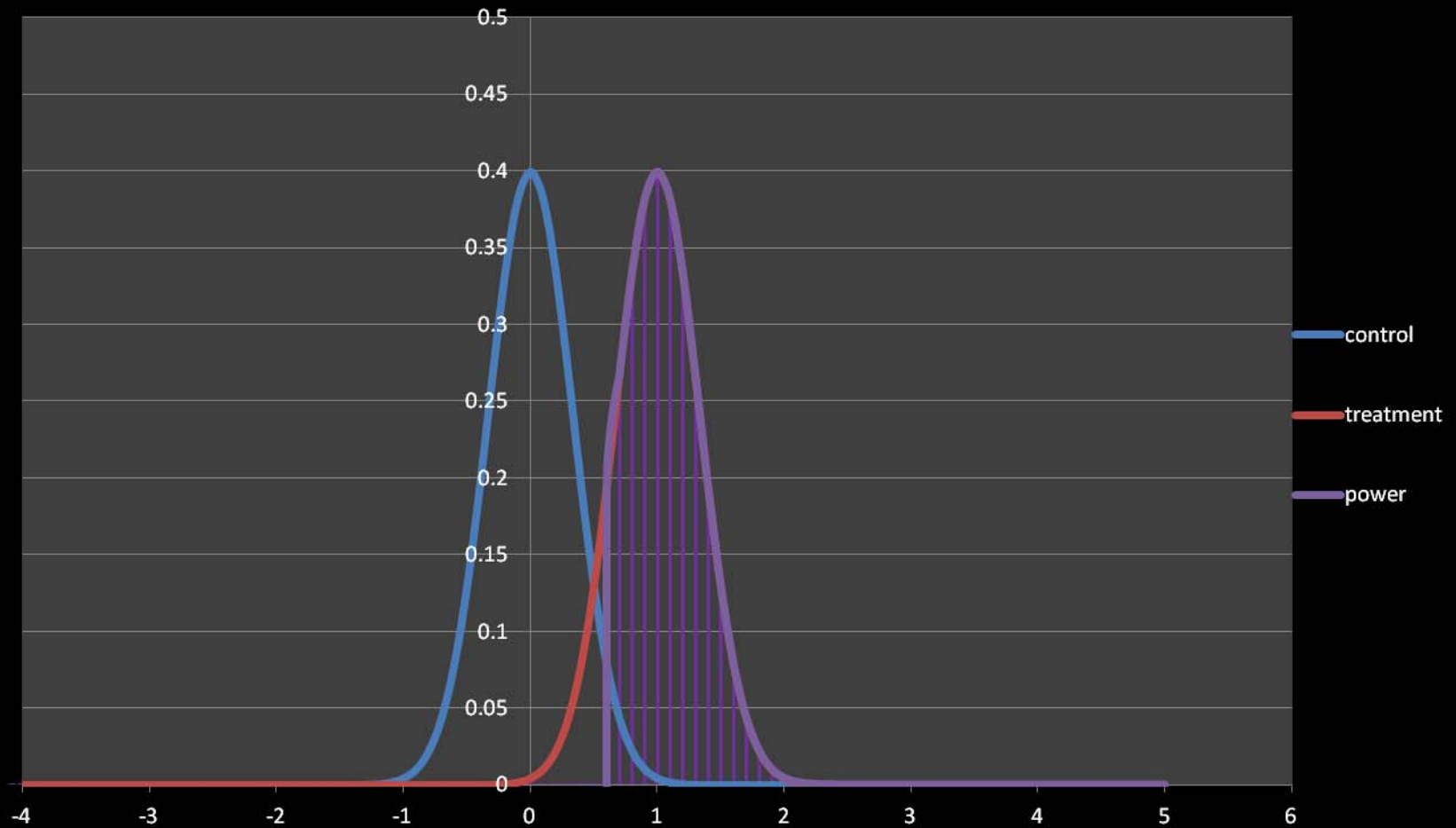
Power: 64%



Power: Sample size = 9



Power: 91%



Sample size

- In this example:
 - a sample size of 9 gave us good power
 - But the effect size we used was very large (1 SD)

Calculating power

- When planning an evaluation, with some preliminary research we can calculate the minimum sample we need to get to.
- A power of 80% tells us that, in 80% of the experiments of this sample size conducted in this population, if H_0 is in fact false (e.g. the treatment effect is not zero), we will be able to reject it.
- The larger the sample, the larger the power.
- Common Power used: 80%, 90%

Clustered design: intuition

- You want to know how close the upcoming national elections will be
- Method 1: Randomly select 50 people from entire Indian population
- Method 2: Randomly select 5 families, and ask ten members of each family their opinion

Clustered design: intuition

- If the response is correlated within a group, you learn less information from measuring multiple people in the group
- It is more informative to measure unrelated people
- Measuring similar people yields less information

Clustered design

- Cluster randomized trials are experiments in which social units or clusters rather than individuals are randomly allocated to intervention groups
- The unit of randomization (e.g. the school) is broader than the unit of analysis (e.g. students)
- That is: randomize at the school level, but use child-level tests as our unit of analysis

Consequences of clustering

- The outcomes for all the individuals within a unit may be correlated
- We call ρ (rho) the correlation between the units within the same cluster

Values of ρ (rho)

- Like percentages, ρ must be between 0 and 1
- When working with clustered designs, a lower ρ is more desirable
- It is sometimes low, 0, .05, .08, but can be high: 0.62

Madagascar Math + Language	0.5
Busia, Kenya Math + Language	0.22
Udaipur, India Math + Language	0.23
Mumbai, India Math + Language	0.29
Vadodara, India Math + Language	0.28
Busia, Kenya Math	0.62

Some examples of sample size

Study	# of interventions (+ Control)	Total Number of Clusters	Total Sample Size
Women's Empowerment	2	Rajasthan: 100 West Bengal: 161	1996 respondents 2813 respondents
Pratham Read India	4	280 villages	17,500 children
Pratham Balsakhi	2	Mumbai: 77 schools Vadodara: 122 schools	10,300 children 12,300 children
Kenya Extra Teacher Program	8	210 schools	10,000 children
Deworming	3	75 schools	30,000 children
Bednets	5	20 health centers	545 women

Implications for design and analysis

- Analysis: The standard errors will need to be adjusted to take into account the fact that the observations within a cluster are correlated.
- Adjustment factor (design effect) for given total sample size, clusters of size m , intra-cluster correlation of r , the size of smallest effect we can detect increases by $\sqrt{1 + \rho * (m - 1)}$ compared to a non-clustered design
- Design: We need to take clustering into account when planning sample size

Implications

- If experimental design is clustered, we now need to consider ρ when choosing a sample size (as well as the other effects)
- It is extremely important to randomize an adequate number of groups
- Often the number of individuals within groups matter less than the total number of groups

MIT OpenCourseWare
<http://ocw.mit.edu>

Resource: Abdul Latif Jameel Poverty Action Lab Executive Training: Evaluating Social Programs
Dr. Rachel Glennerster, Prof. Abhijit Banerjee, Prof. Esther Duflo

The following may not correspond to a particular course on MIT OpenCourseWare, but has been provided by the author as an individual learning resource.

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.