LECTURE 4: POWER AND SAMPLE SIZE FOR CLUSTERED RCTS

Karen Macours
Paris School of Economics

Overview

- Sampling and a common confusion
  - Internal and external validity
- Clusters and stratification
- Power calculations and the sample size question
- Some rules of thumb – main ideas

Sampling and a common confusion

- Think of a 2-step process:
  - In the first step, a random sample of participants is selected from a defined population

    => This gives a representative subset of the population
  - In the second step, part of this sample is randomly assigned to treatment, and the other part to the control

    => This gives me my much desired counterfactual

External and internal validity

- First Step => External validity
  - The sample is representative of the total population
  - The results in the sample represent the results in the population
  - We can apply the lessons to the whole population
- Second Step => Internal validity
  - The estimated effect of the intervention/program on the evaluated population reflects the *real* impact on that population
  - i.e. the intervention and comparison groups are comparable

External and internal validity

- An evaluation can have internal validity without external validity
  - Example: a randomized evaluation of encouraging women to change dietary diversity in urban area may not tell you much about impact of a similar program in rural areas
- An evaluation without internal validity, can't have external validity
  - If you don't know whether a program works in one place, then you have learnt nothing about whether it works elsewhere.

External Validity and Sampling

- Remember we want to compare means of treatment and control
- But we do not observe the whole population (typically), but only a sample
- We hence obtain an estimate of the population means by computing the average in the treatment and control sample
  - But if we have very few observations, the averages are imprecise. When we see a difference in sample averages, we do not know whether it comes from the effect of the treatment or from something else

    => calculate confidence intervals
- Given that we know this, how should we sample?

Sample selection for impact evaluation

- Population based representative surveys:
  - Sample representative of whole population
  - Good for learning about the population
  - Not always most efficient for impact evaluation
- Sampling for Impact evaluation
  - Balance between treatment and control groups
  - Power => statistical inference for groups of interest
- Survey budget as major consideration
  - In practice, sample size is often set by budget
  - Concentrate sample on key populations to increase power

Purposive Sampling

- Risk: We will systematically bias our sample, so results don't generalize to the rest of the population or other sub-groups
- Trade off between power within population of interest and population representation
- Results are internally valid, but not generalizable.

Overview

- Sampling and a common confusion
  - Internal and external validity
- Clusters and stratification
- Power calculations and the sample size question
- Some rules of thumb – main ideas

Sampling frame

- Simple Sampling
  - almost never practical unless universe of interest is geographically concentrated
- Cluster Sampling
  - randomly chose clusters and then randomly chose units within the cluster. Effective sample size is less than actual number of observations. This is the design or cluster effect
  - The design effect implies that, for a given sized sample, the variance increases [1 + ρ(E-1)]
    - where E is the number of elements in each cluster and
    - ρ is the intra-class correlation, a measure of how much the observations with in a cluster resemble each other.

More on clustered s.e.

- What does clustering mean for our estimates?
  - OLS assumes that s.e. are i.i.d. "independently and identically distributed"
  - In this case: observations are not independent from each other
  - Need to adjust the variance-covariance matrix to account for the intercluster correlation
  - This will generally lead to higher standard errors (less precision) – compared to i.i.d.
  - In stata: "cluster" option: calculates generalized "White" covariance matrix

Stratification

- Objective: balancing your sample when you have a small sample
  - on variables that could have important impact on outcome variable (bit of a guess)
  - Stratify on subgroups that you are particularly interested in (where may think impact of program may be different)

- What is it:
  - dividing the sample into different subgroups
  - selecting treatment and control from each subgroup

- What happens if you don't stratify?

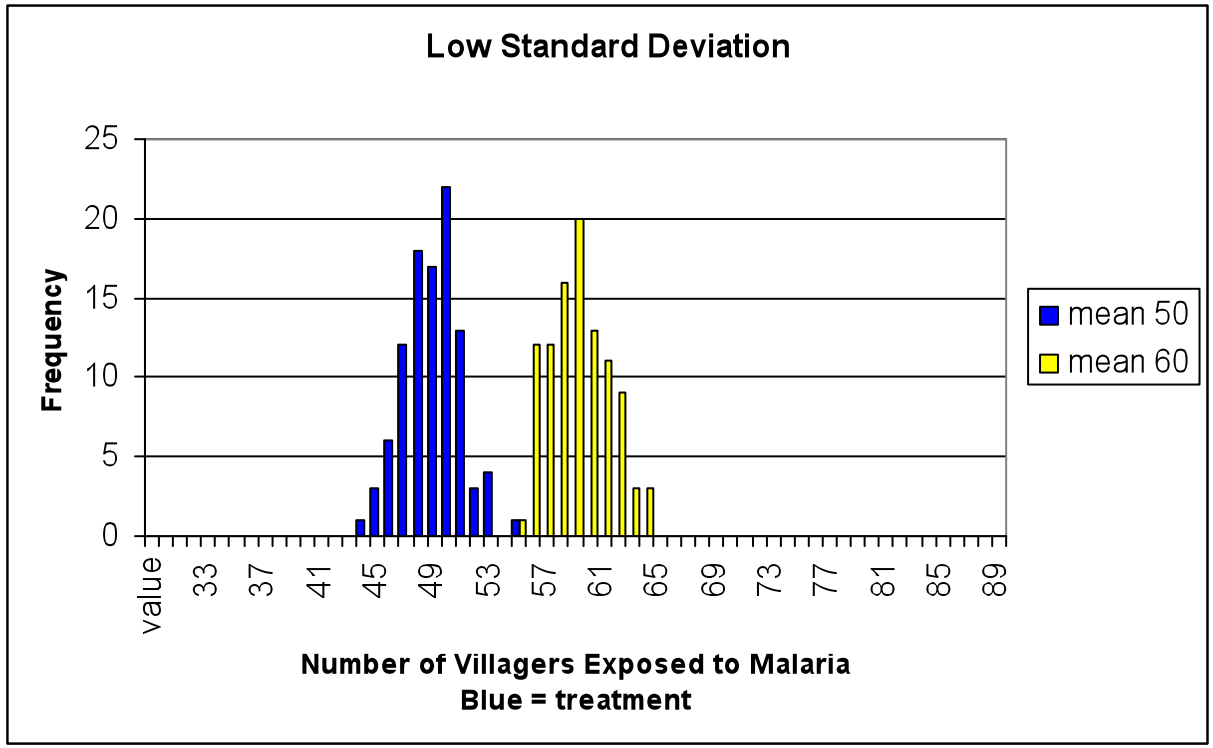
Overview

- Sampling and a common confusion
  - Internal and external validity
- Clusters and stratification
- Power calculations and the sample size question
- Some rules of thumb – main ideas

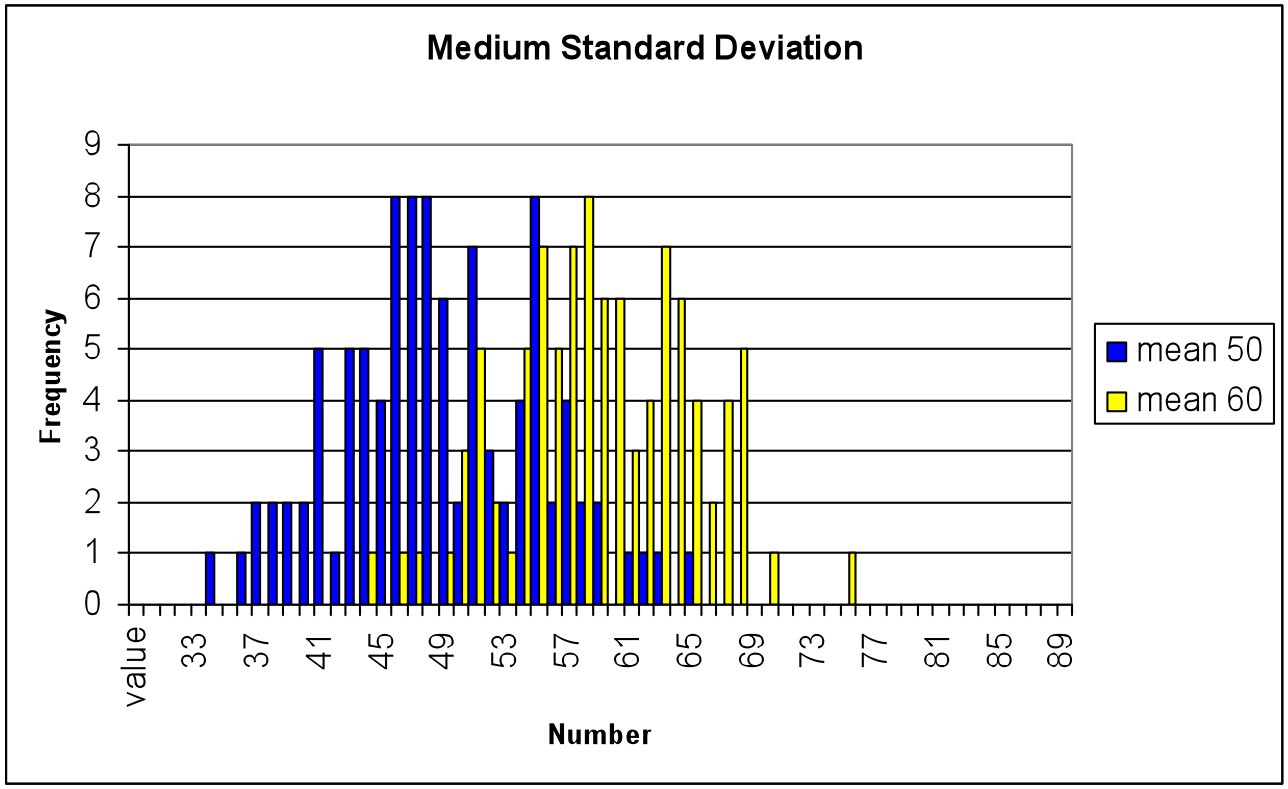
Sample size question

- 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?
- All of these questions can be answered via **POWER**

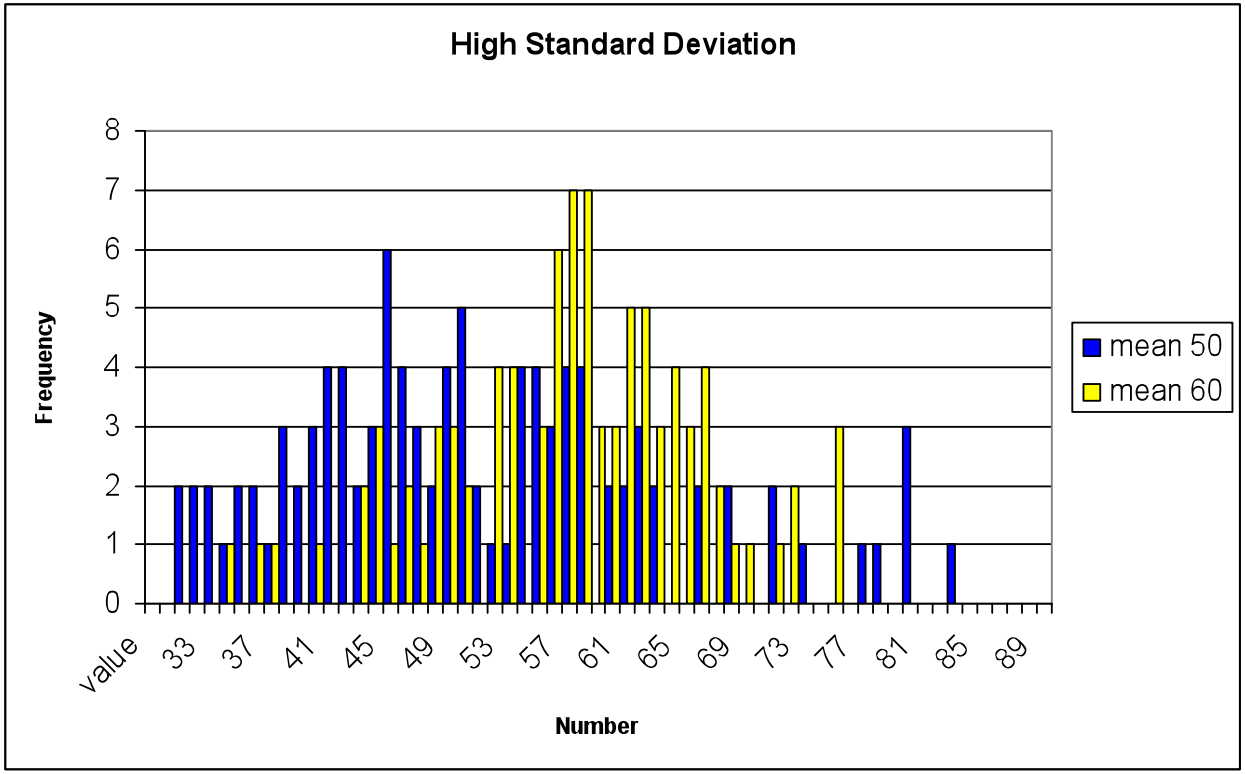
Why do we care about noise?
Precise outcomes



Why do we care about noise? Some noise



Why do we care about noise?
Very noisy



Two Types of Errors

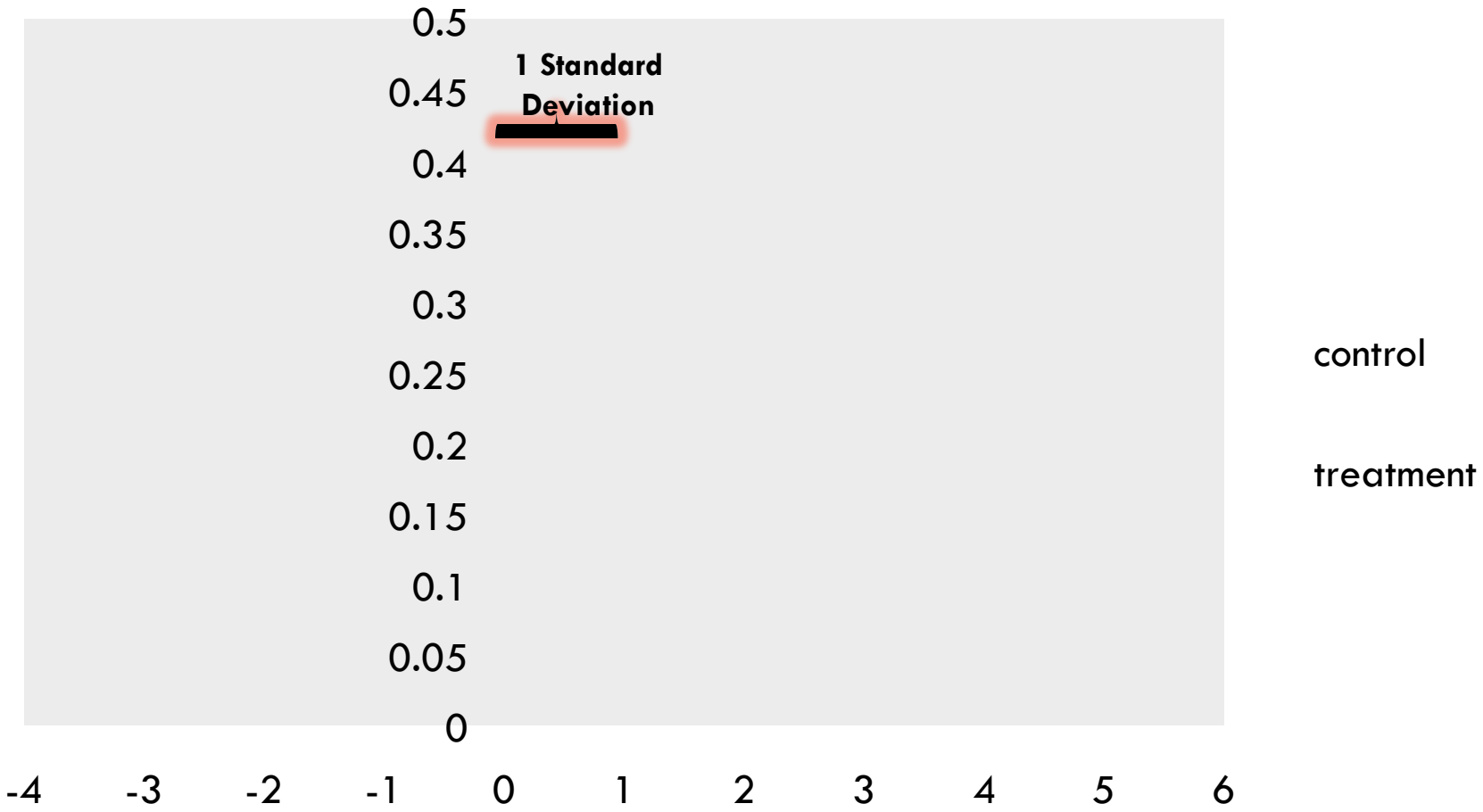
- Type I error : Conclude that there is an effect, when in fact there is no effect.
  - The level of your test is the *probability that you will falsely conclude that the program has an effect, when in fact it does not.*
  - So with a level of 5%, you can be 95% confident in the validity of your conclusion that the program had an effect.

- Type II of error: you fail to reject that the program had no effect, when it fact it does have an effect.
  - The Power of a test is the probability of finding a significant effect in the RCT

Null Hypothesis: assume zero impact

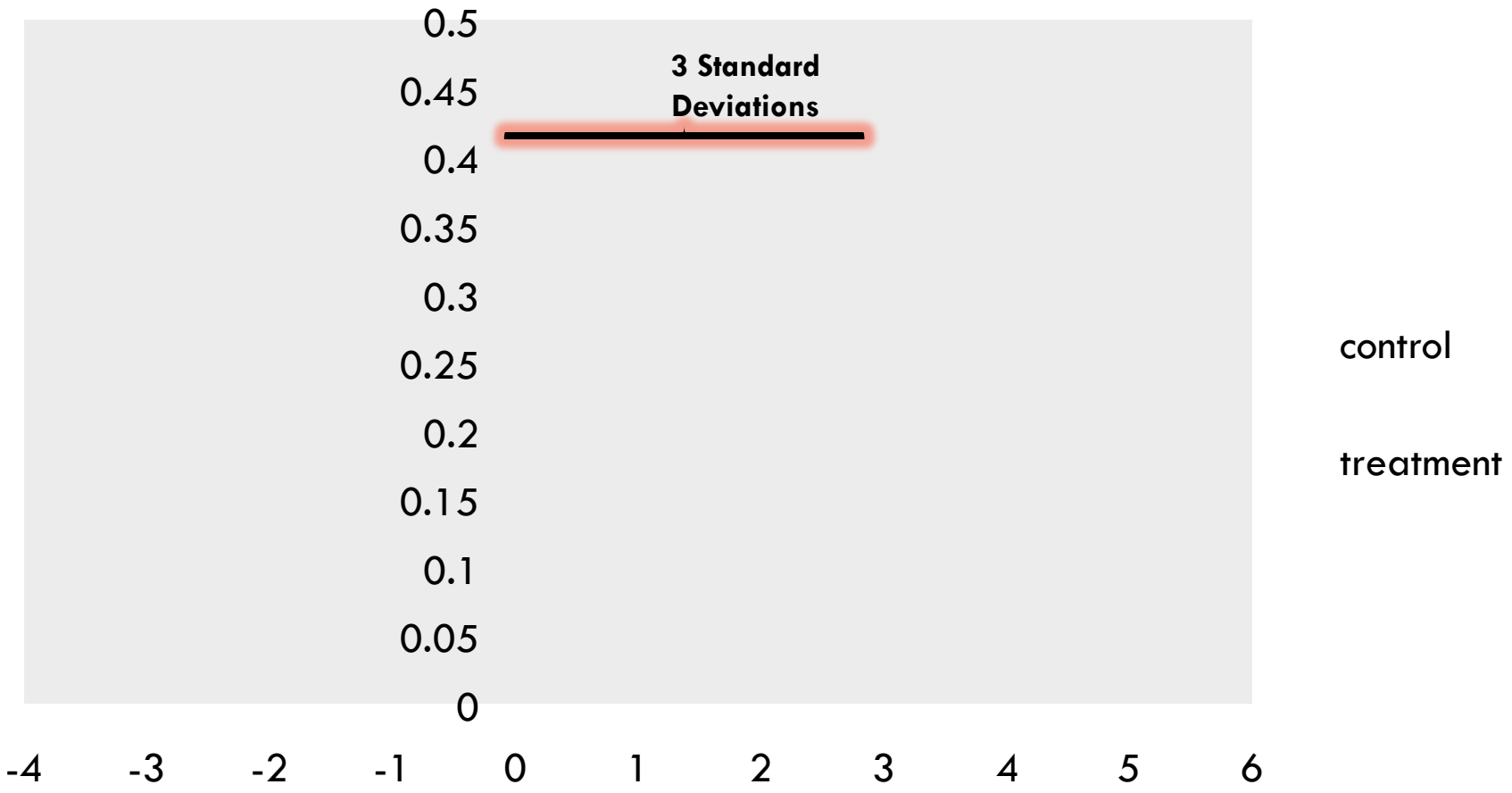
"Impact = 0" There's a sampling distribution around that.

Effect Size: 1 “standard deviation”



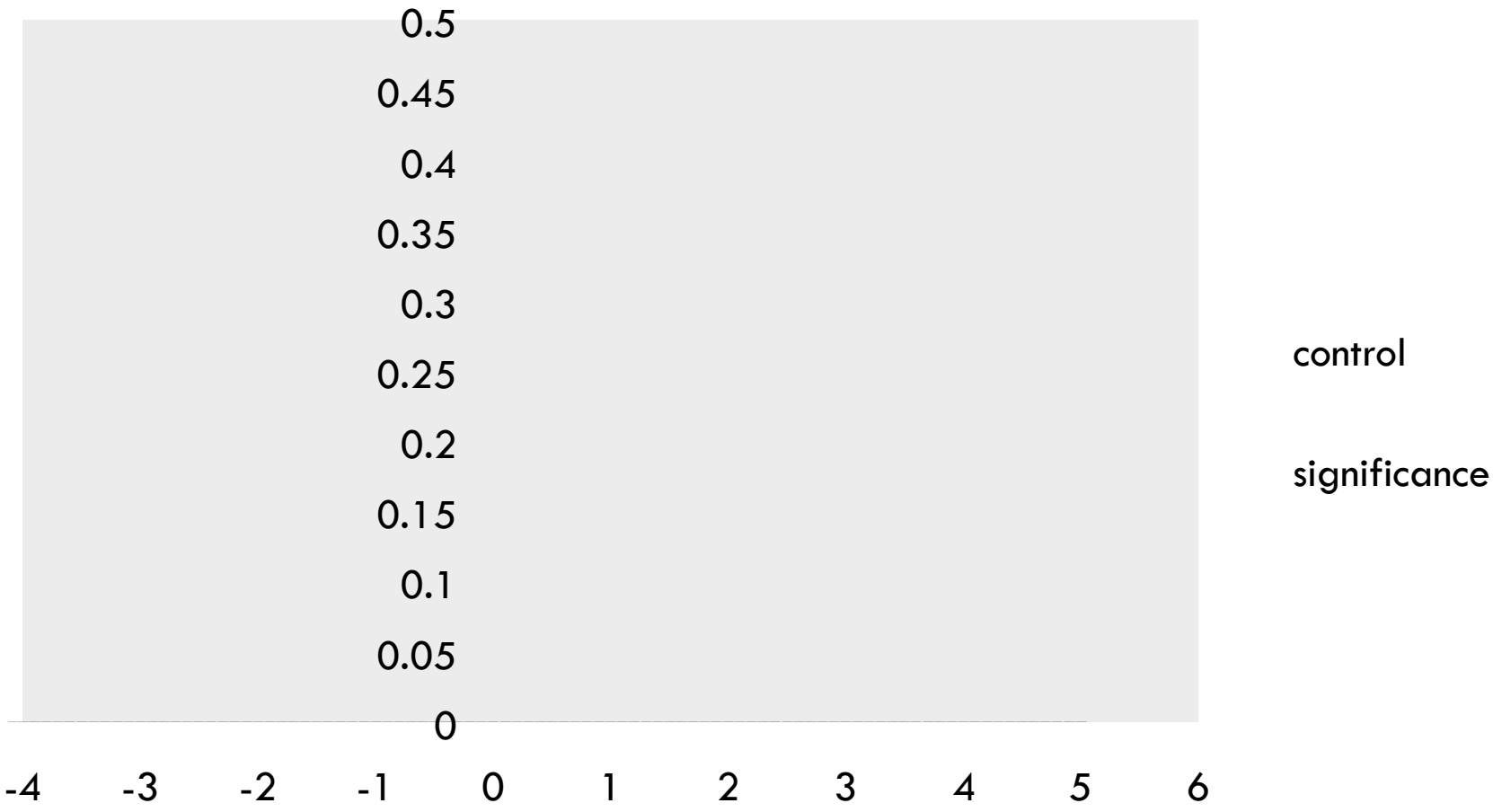
And there’s a new sampling distribution around that

Effect Size: 3 standard deviations

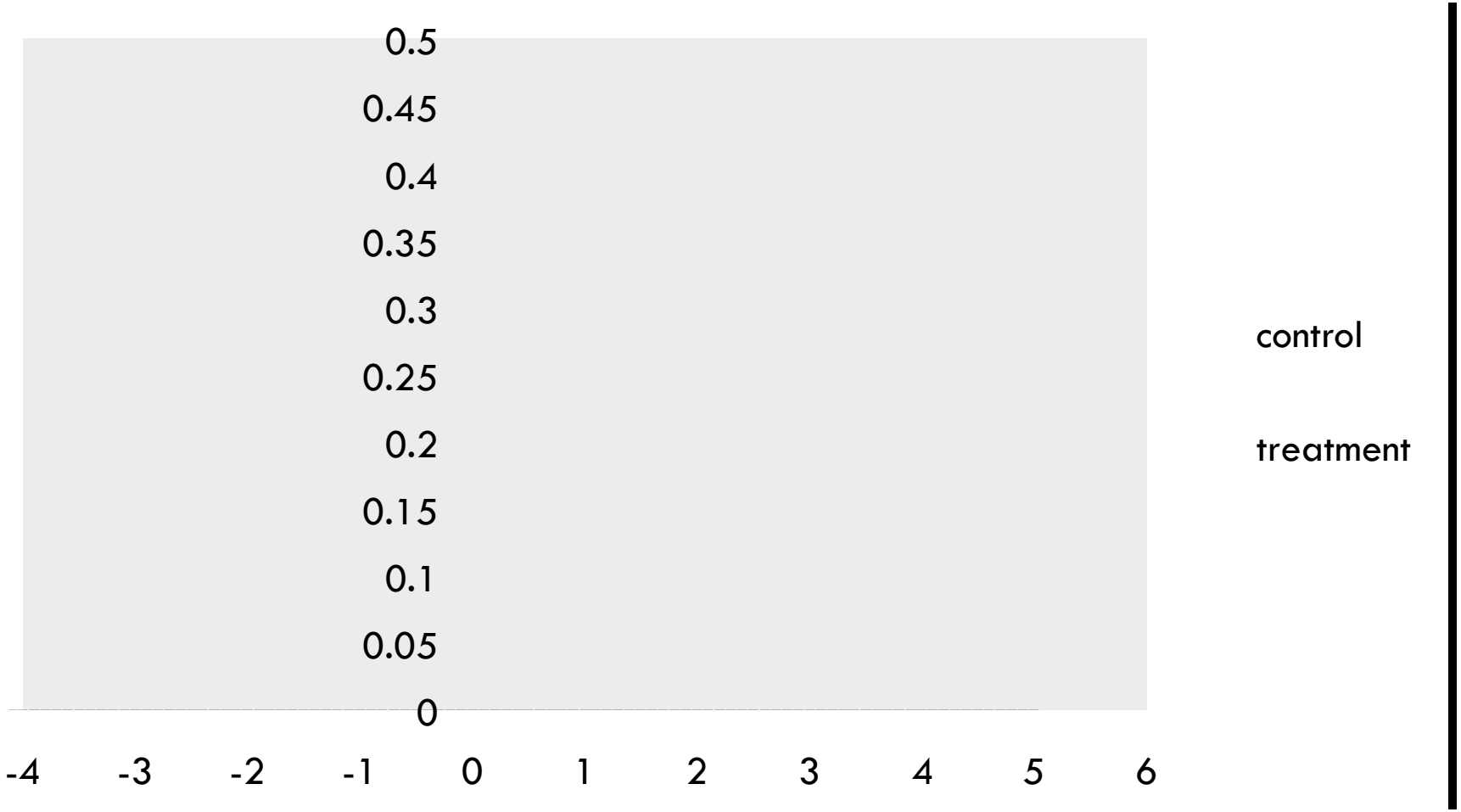


The less overlap the better…

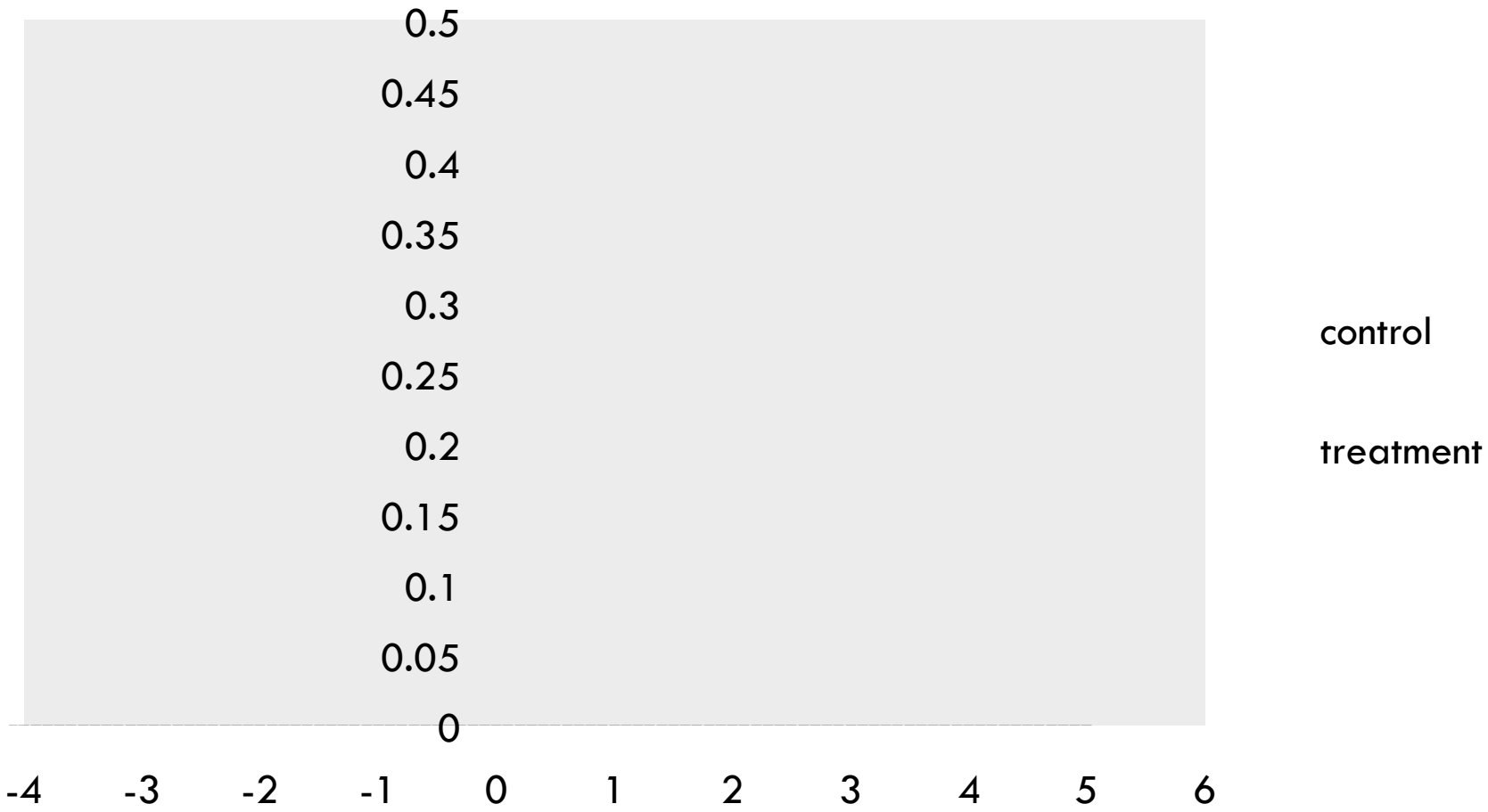
Sign. level: reject H_0 in critical region



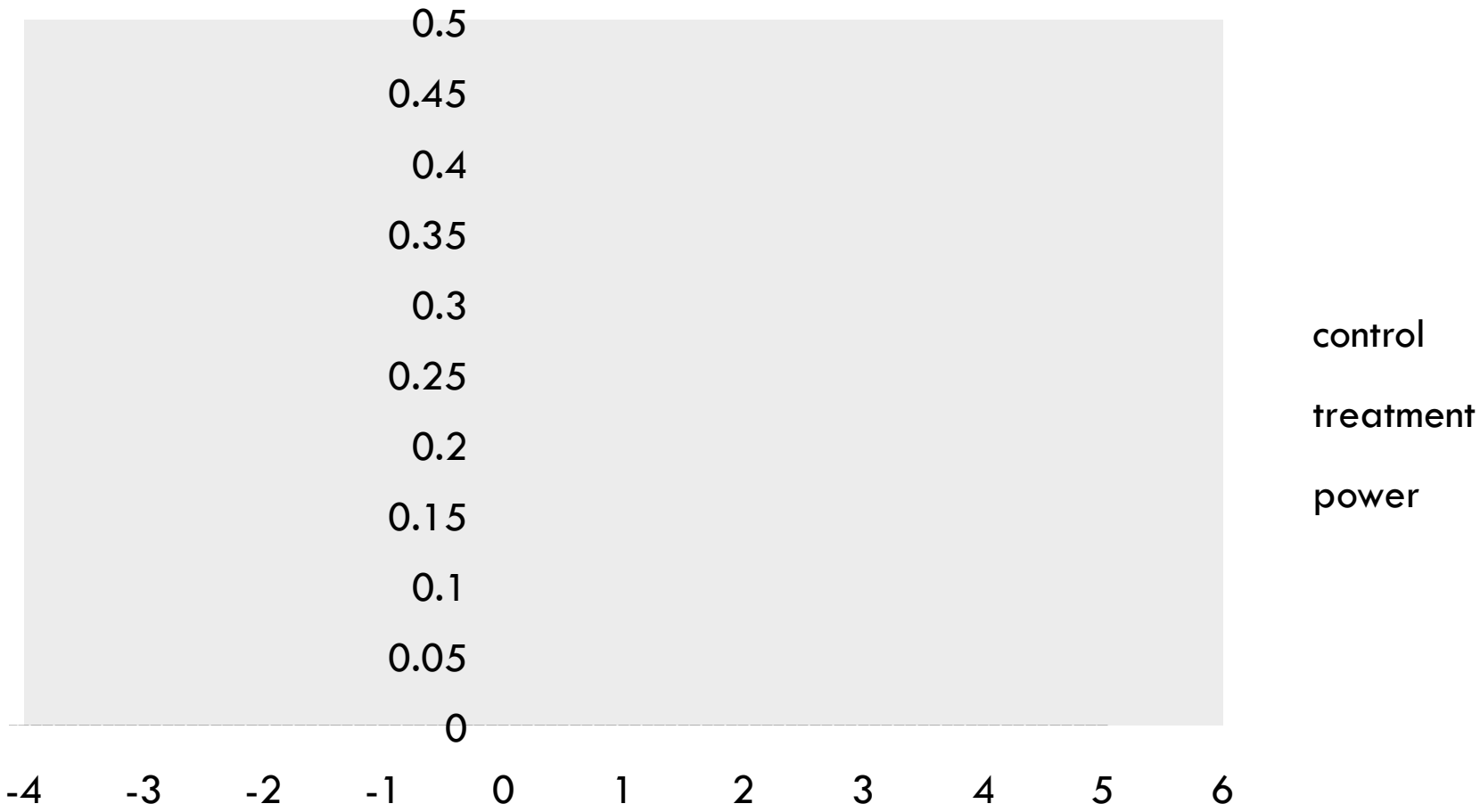
True effect is 1 SD



Power: when is H_0 rejected?

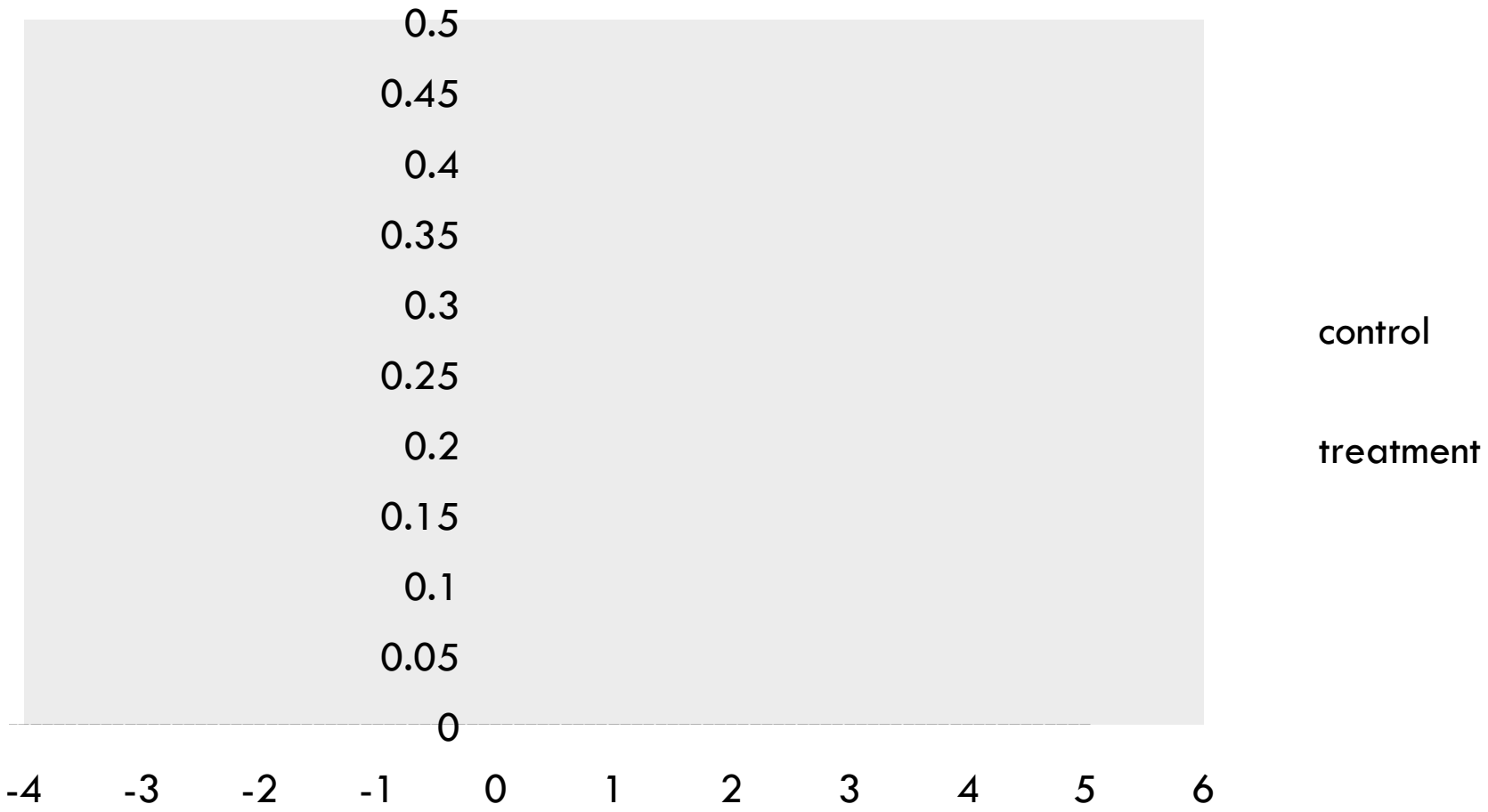


If the true impact was 1SD…

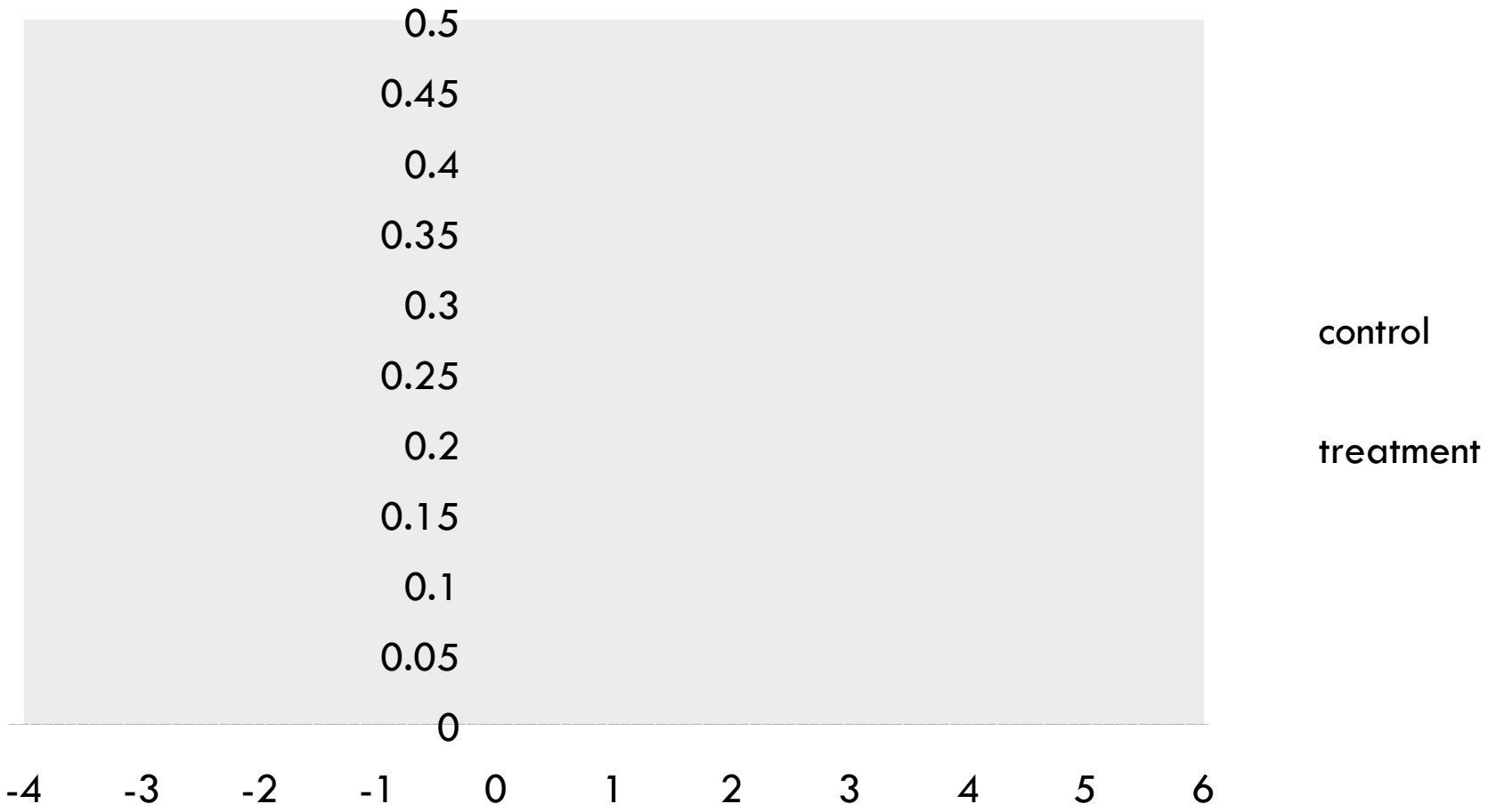


The Null Hypothesis would be rejected only 26% of the time

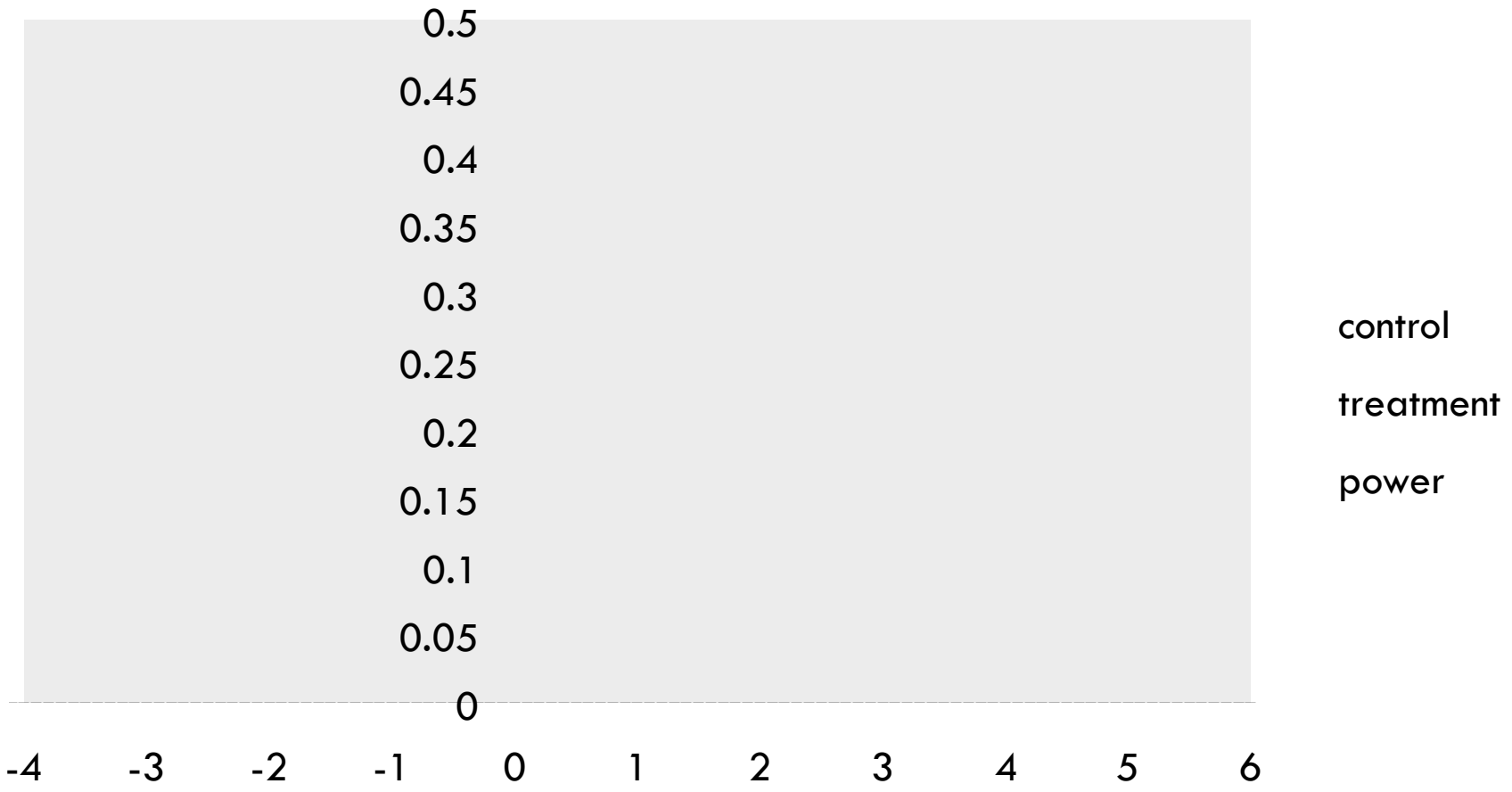
Power: if we change the effect size?



Power: assume effect size = 3 SDs



Power: 91%



The Null Hypothesis would be rejected 91% of the time

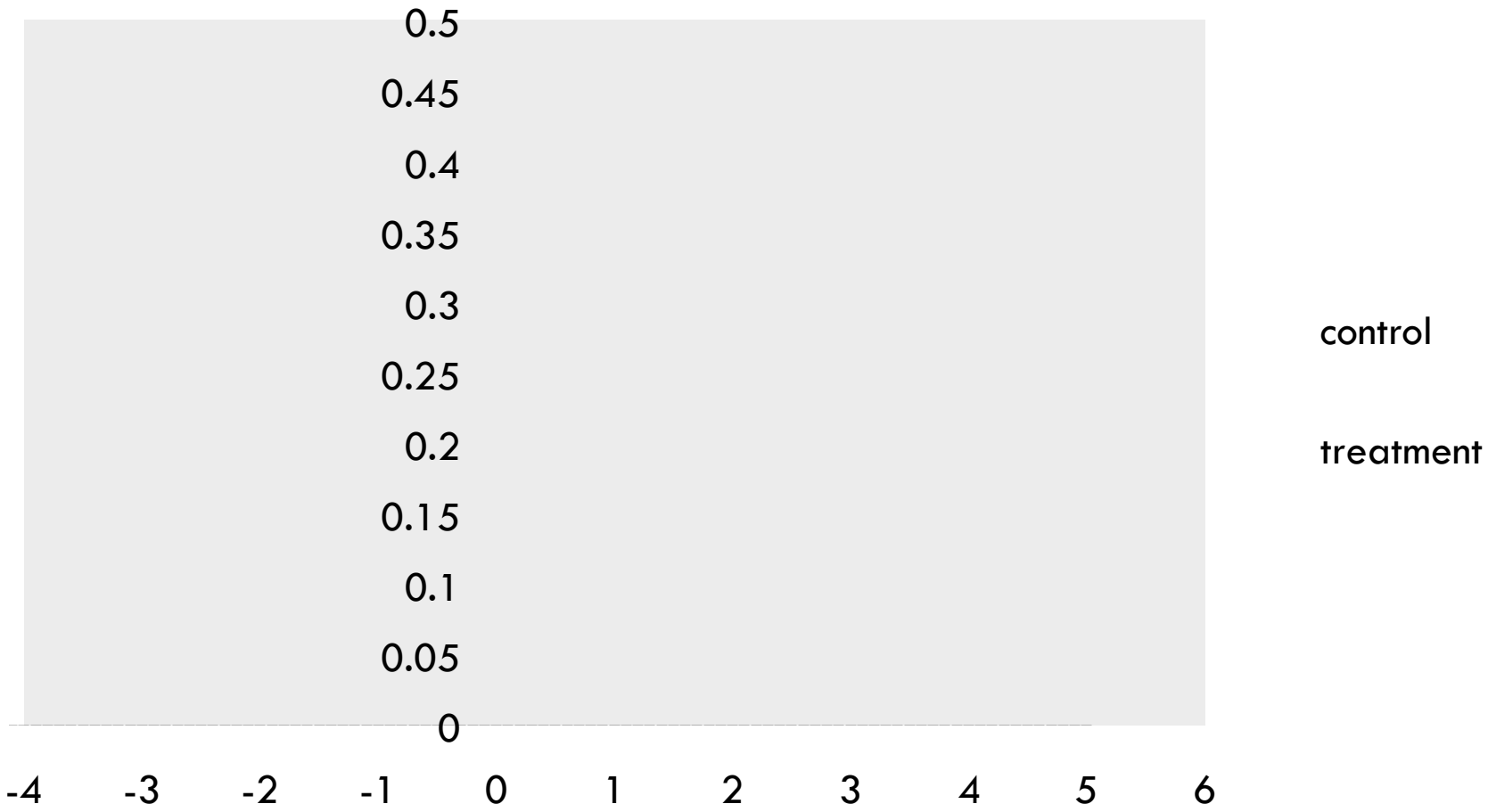
How to Determine Effect Size

- What is the smallest effect that should justify the program to be adopted (in terms of cost-benefit)?
  - Sets minimum effect size we would want to be able to test for
- Common danger: use an effect size that is too optimistic ‣too small of sample size
- How large an effect you can detect with a given sample depends on how variable the outcomes is.
  - Example: If all children have very similar diarrhea prevalence without a program, a very small impact will be easy to detect

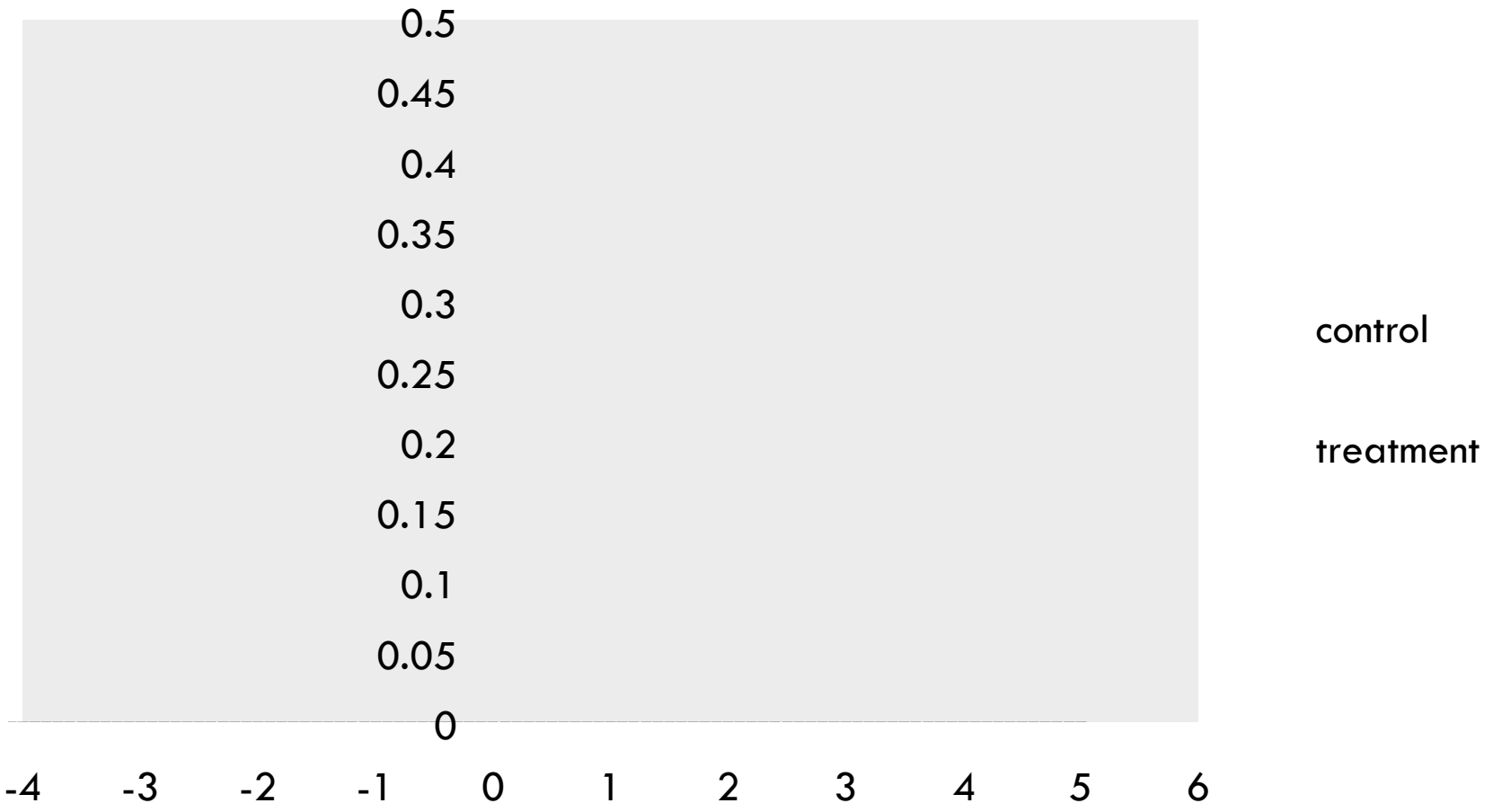
Power calculations to estimate sample size

- What is the sample size needed to be able to find a difference in means at a given statistical significance?
  - Need ideal of what difference is a plausible expectation for the intervention.
    - Minimum detectable effect (MDE), usually determined via policy and cost-effectiveness considerations.
  - Fixing the confidence level, we observe two things when increasing sample size:
    - the rejection region gets larger and
    - the power increases

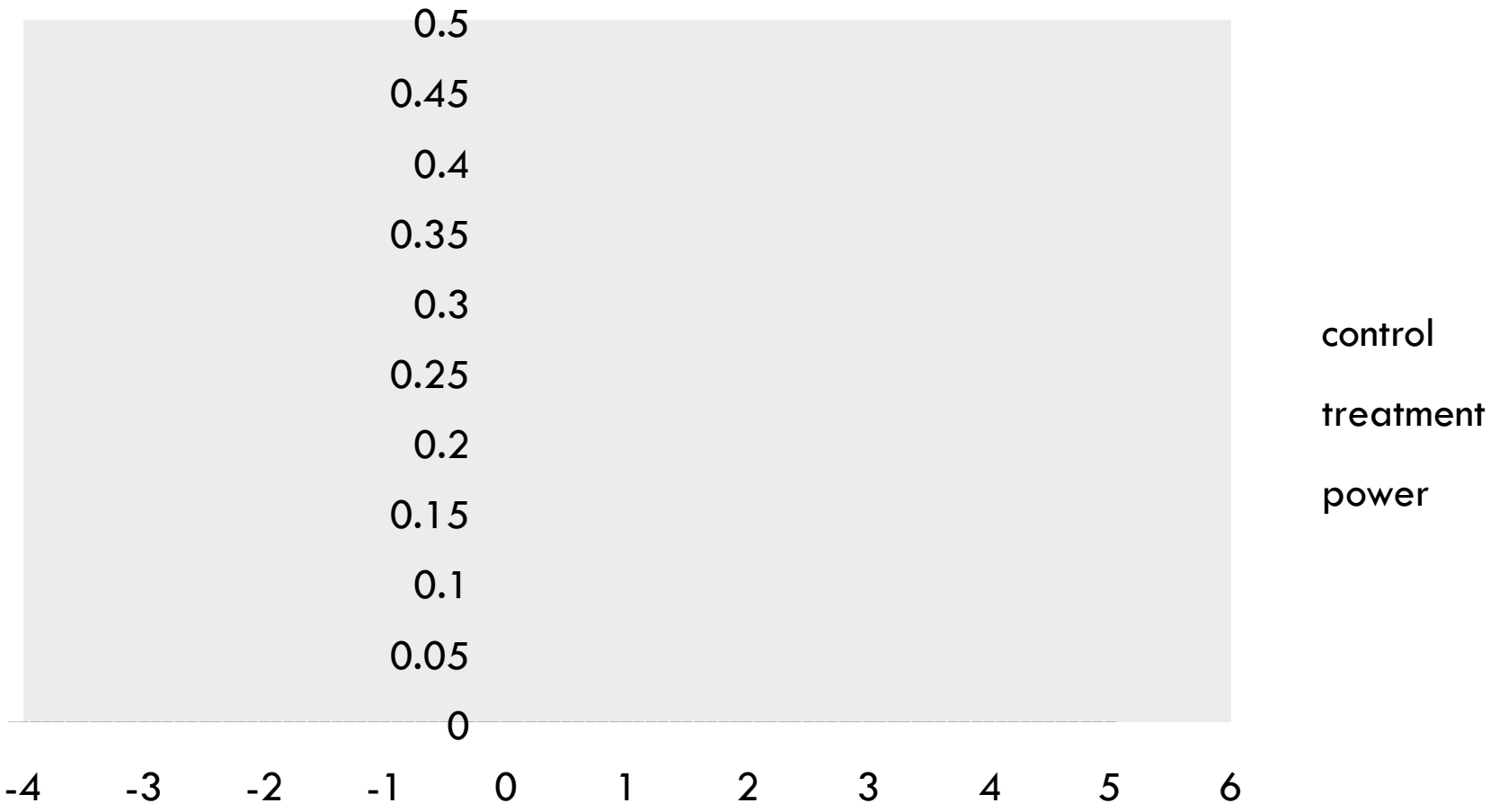
Power: Effect size = 1SD, Small sample



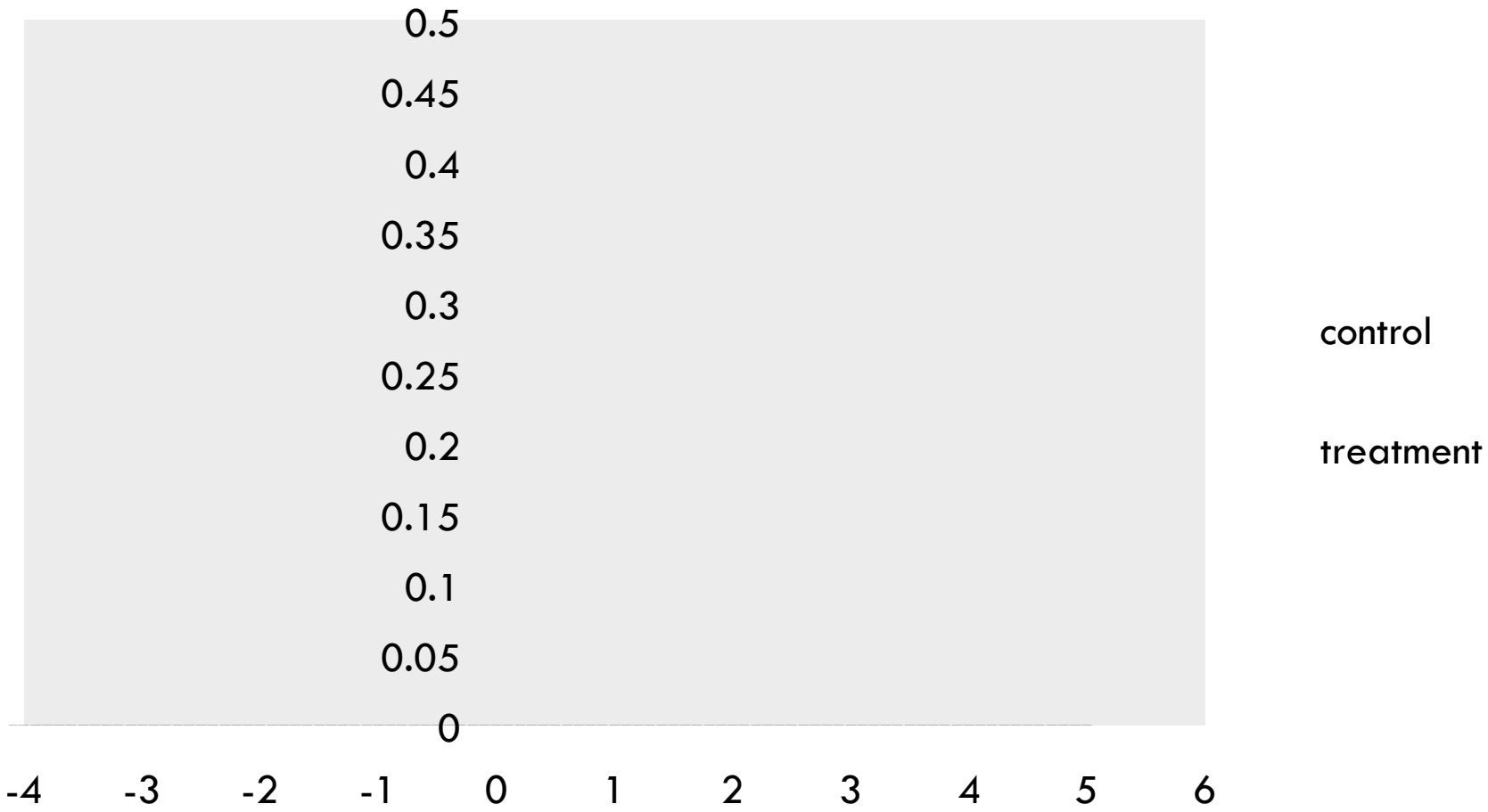
Power: Let's increase sample size



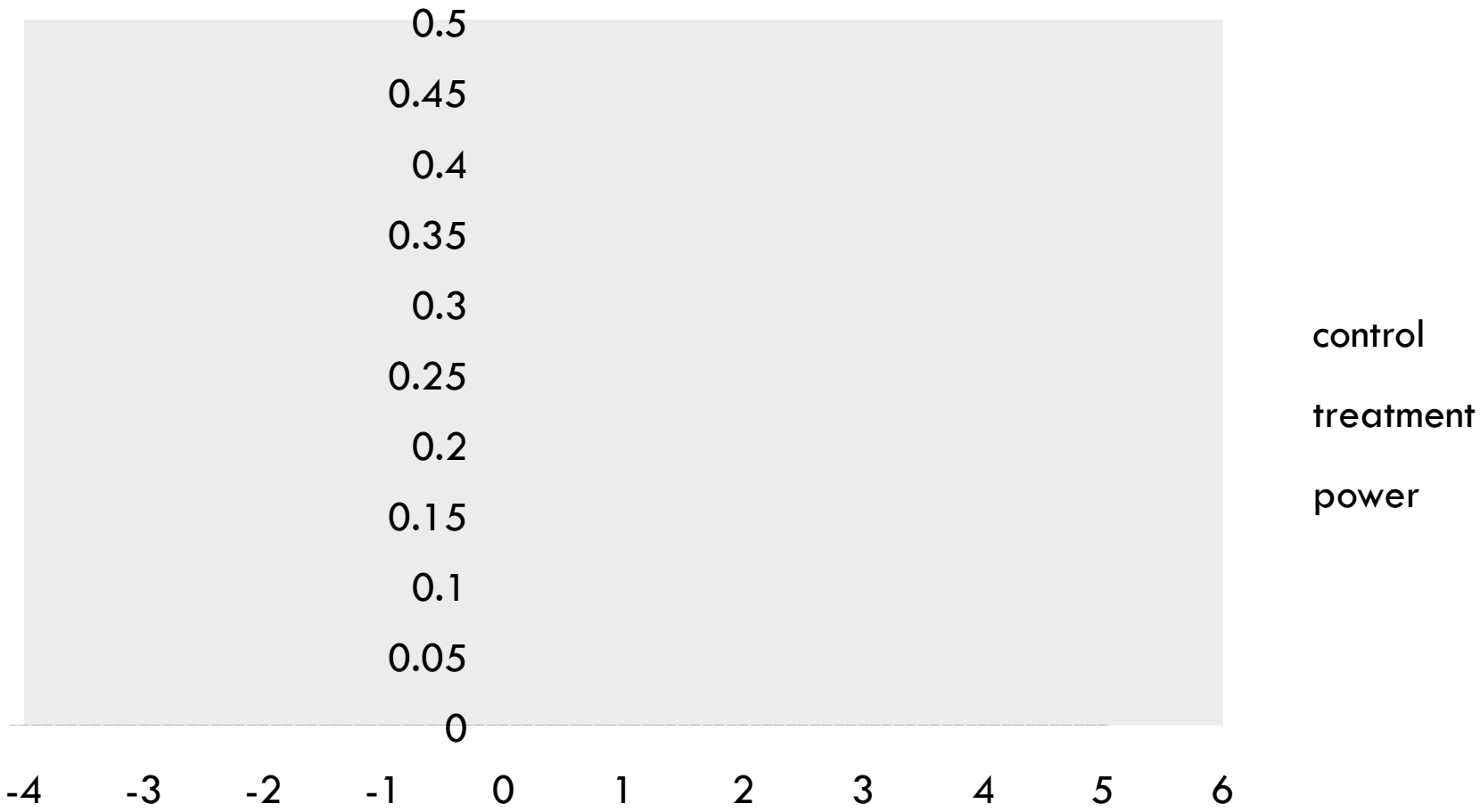
Power: 64%


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
control
treatment
power

Power: Increase sample size further



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
control
treatment
power

More formally (1)

- The variance of the point estimate in OLS:

$$\frac{1}{P(1-P)}\frac{\sigma^2}{N}$$

- Power κ: fraction of area of the distribution of that estimate that falls to the right of the critical value. To achieve power κ

$$\beta > (t_{1-\kappa} + t_{\alpha})SE(\hat{\beta})$$

More formally (2)

- The minimum detectable effect size

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

- Points to relationship between MDE, number of observations, Proportions in treatment and control, noise, power and significance

Power calculations in practice

- Factors affecting sample size?
  - Desired power
  - Desired significance level
  - Expected magnitude of impact (or minimum magnitude at which we want to find an impact)
  - Standard error (how noisy is the data)
    - Control variables & baseline: can I take out some noise?
  - Proportion of observations in 2 groups

- What if design is a bit more complicated?
  - Intra-cluster correlation
  - Number of clusters (may be limited)
  - Inter-temporal correlation and number of rounds of data in case of panel
  - Might there be partial compliance?
  - Might there be attrition (in program, but also in survey)

Where do we get all this stuff? (1)

- We need to set :
  - Desired significance level (often used 0.05)
  - Desired power (often used levels 0.8 and 0.9)
  - Minimum magnitude of impact that we would want to be able to differentiate from 0 (MDE).
    - Can be an absolute level, informed by
      - cost-benefit calculation
      - previous evaluations
      - Level & variation in the data
    - Can also be a standardized effect size (the effect size divided by the standard deviation of the outcome)
      - Common effect sizes: .20 (small); .50 (medium); .80 (large)

Where do we get all this stuff? (2)

- We need to make an informed guess about:
  - Standard deviation (or else think about magnitude of impact in terms of s.d.)
  - Intra-cluster correlation
  - Inter-temporal correlation and number of rounds of data in case of panel

  => Use data from the population of interest (or of "similar" populations)

- If design permits, we can try different combinations of
  - # clusters
  - Proportion of observation in treatment and control

What if we have more than 1 treatment?

- Depends on unit of randomization of each treatment
  - Units could differ, in particular when not testing differences between interventions, but rather possible complementarities
- When comparing treatments -> remember you might want to pick up difference in effect size, which could (or could not) be larger than difference between treatment and control
- Think of potential differences in compliance for different treatments

Power calculations in practice

- Many analytical statistical results especially when one can vary cluster numbers and cluster sizes
- May use simulations in Stata or specific power calculation software (Optimal design).
- http://sitemaker.umich.edu/group-based/optimal_design_software

Overview

- Sampling and a common confusion
  - Internal and external validity
- Clusters and stratification
- Power calculations and the sample size question
- Some rules of thumb – main ideas

Rules of thumb (1)

1. A larger sample is needed to detect differences between two variants of a program than between the program and the comparison group.

2. For a given sample size, the highest power is achieved when half the sample is allocated to treatment and half to comparison.

Rules of thumb (2)

3. The more measurements are taken, the higher the power. In particular, if there is a baseline and endline rather than just an endline, you have more power.
4. The lower compliance, the lower the power. The higher the attrition, the lower the power.
5. For a given sample size, we have less power if randomization is at the group level than at the individual level.

Recap on Clustered design

- Intuition of clustered design:
  - if all observations in a cluster are exactly the same, the number of clusters = number of observations

  => The more similar the observations in a cluster, the lower the additional variance we obtain from adding sample size within the cluster
- Number of clusters improves precision and is important especially in randomized designs.
- Number of observations per group might not matter as much as number of groups

  => Make sure to get enough clusters
- Not strictly necessary that treatment and control are equal in size or number of clusters.

How to increase power?

- The looser the level of significance we impose
  => the more likely we are to reject the null, i.e. the higher the power:
  => but also the more likely we are to make false positive (type II) errors.
- Higher MDE => higher power.
- The lower the variance of the underlying population
  => the lower the variance of the estimated effect size and the higher the power.
- The larger the sample size
  => the lower the variance of our estimate effect and the higher the power.
- The more evenly the sample is distributed between T and C
  => the higher the power.
- Individual-level randomization is more powerful than group-level randomization given the same sample size.
- More outcomes correlated within groups in a group-level randomization => less power.

Common tradeoffs

- Few clusters with many observations? Or many clusters with few observations?
- Answer one question really well? Or many questions with less accuracy?
- Large sample size with possible attrition? Or small sample size that we track very closely?
- How do we allocate our sample to each group?

Questions?