EGAP Learning Days: Power Analysis

Gareth Nellis

University of California, Berkeley
Postdoctoral Fellow, Evidence in Governance and Politics

February, 2017

Preliminaries: Average Treatment Effect

Question: How do we calculate the estimated average treatment effect?

Preliminaries: (Estimated) Average Treatment Effect

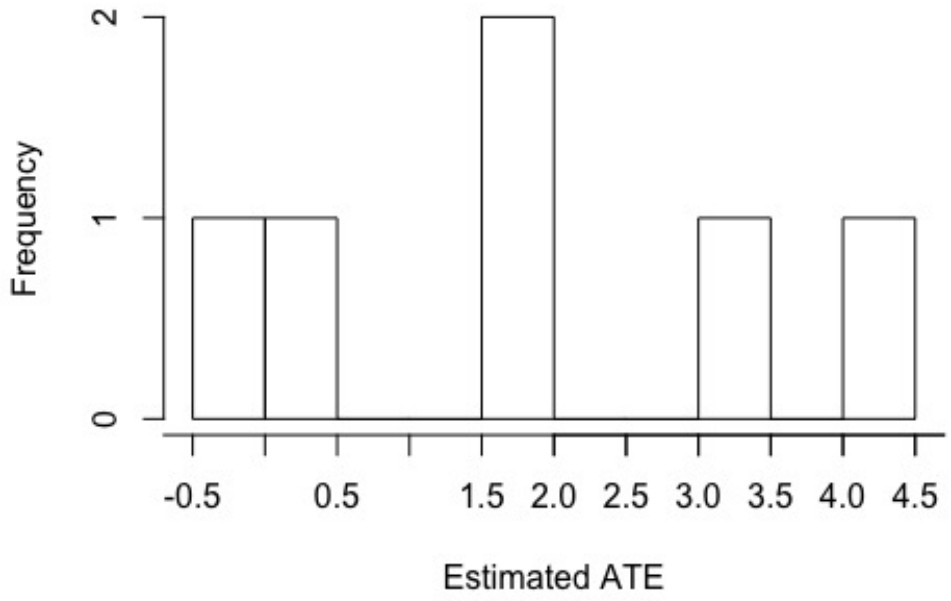
- There is a true average treatment effect in the world
- We try to estimate it, usually using *a single experiment*
- Estimated ATE = (Average outcomes of treatment units) - (Average outcomes of control units)
- If we repeated the experiment again and again, for all possible ways treatment could be assigned, the average of all those estimated ATEs would converge on the true ATE (unbiasedness)
- But we only get to run a single experiment & the estimated ATE from that experiment may be high or may be low

Preliminaries: What is a Sampling Distribution?

Definition: the distribution of estimated average treatment effects for all possible treatment assignments

Sampling Distribution

Say we have an experiment in which 2 of 4 units are randomly assigned to treatment



Schedule of potential outcomes:

Unit	$Y_i(1)$	$Y_i(0)$
a	8	4
b	6	3
c	5	2
d	1	3

$E[Y_i(1) - Y_i(0)] = 2.0$

$\widehat{ATE} =$

$\{-0.5, 0.5, 2.0, 2.0, 3.5, 4.5\}$

Let's Do the Calculation!

	T	C
Unit a	8	
Unit b	6	
Unit c		2
Unit d		3

Diff-in-means = [(8+6)/2] – [(2+3)/2] = 4.5

	T	C
Unit a		4
Unit b		3
Unit c	5	
Unit d	1	

Diff-in-means = [(5+1)/2] – [(4+3)/2] = -0.5

	T	C
Unit a	8	
Unit b		3
Unit c	5	
Unit d		3

Diff-in-means = [(8+5)/2] – [(3+3)/2] = 3.5

	T	C
Unit a		4
Unit b	6	
Unit c		2
Unit d	1	

Diff-in-means = [(6+1)/2] – [(4+2)/2] = 0.5

	T	C
Unit a	8	
Unit b		3
Unit c		2
Unit d	1	

Diff-in-means = [(8+1)/2] – [(3+2)/2] = 2

	T	C
Unit a		4
Unit b	6	
Unit c	5	
Unit d		3

Diff-in-means = [(6+5)/2] – [(4+3)/2] = 2

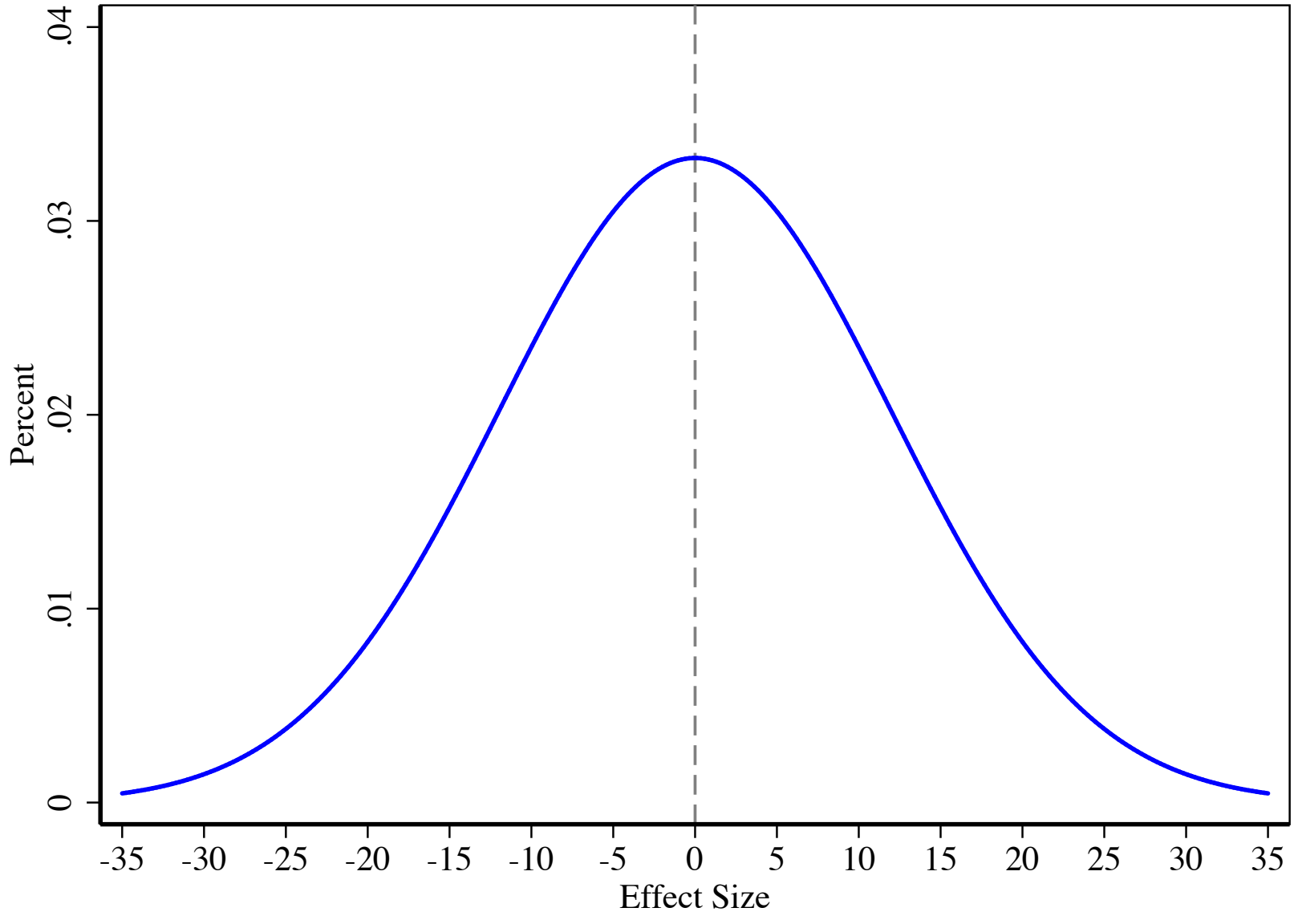
Preliminaries: What is a Variance and a Standard Deviation?

- A measure of the dispersion or spread of a statistic
- Variance: mean-square deviation from average of a variable
- $Var(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- Standard deviation is the square root of the variance
- $SD_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$
- Example: Age

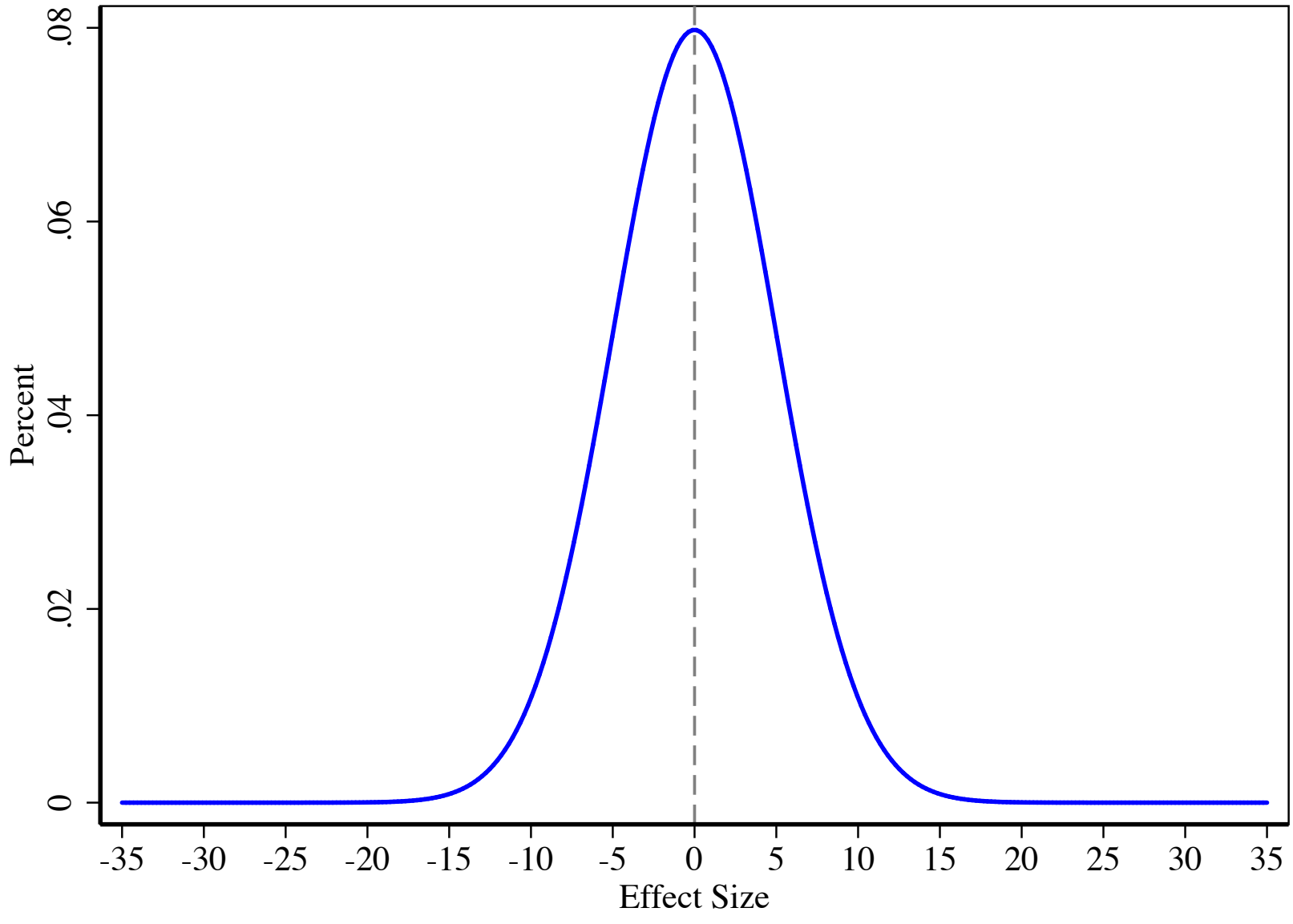
Preliminaries: What is a Standard Error?

- Simple! *The standard deviation of a sampling distribution*
- A measure of sampling variability
- Bigger standard error means that our estimate is more uncertain
- For precise estimates, we need the standard error to be small relative to the treatment effect we're trying to estimate

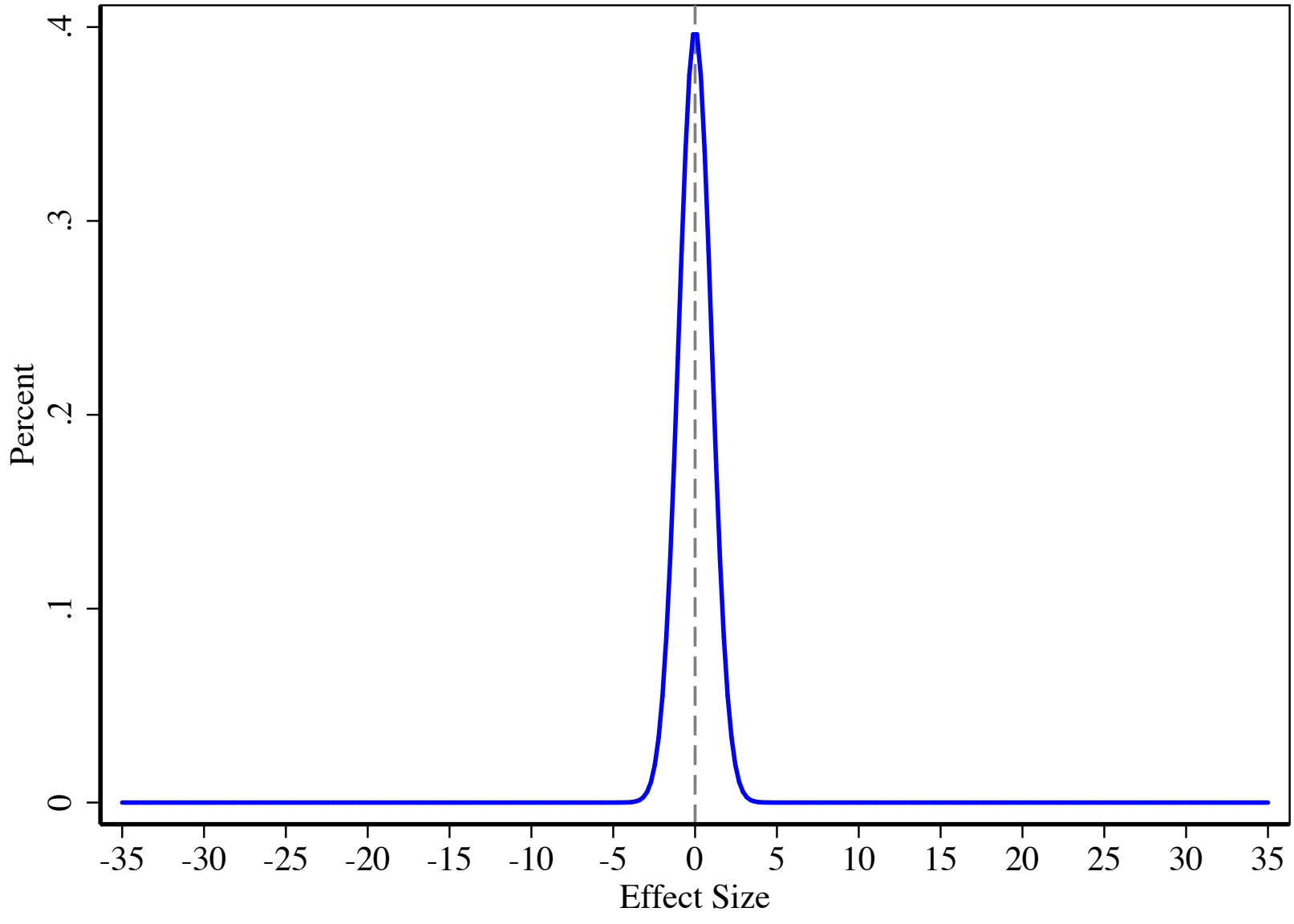
Sampling Distribution: Large-Sample Example



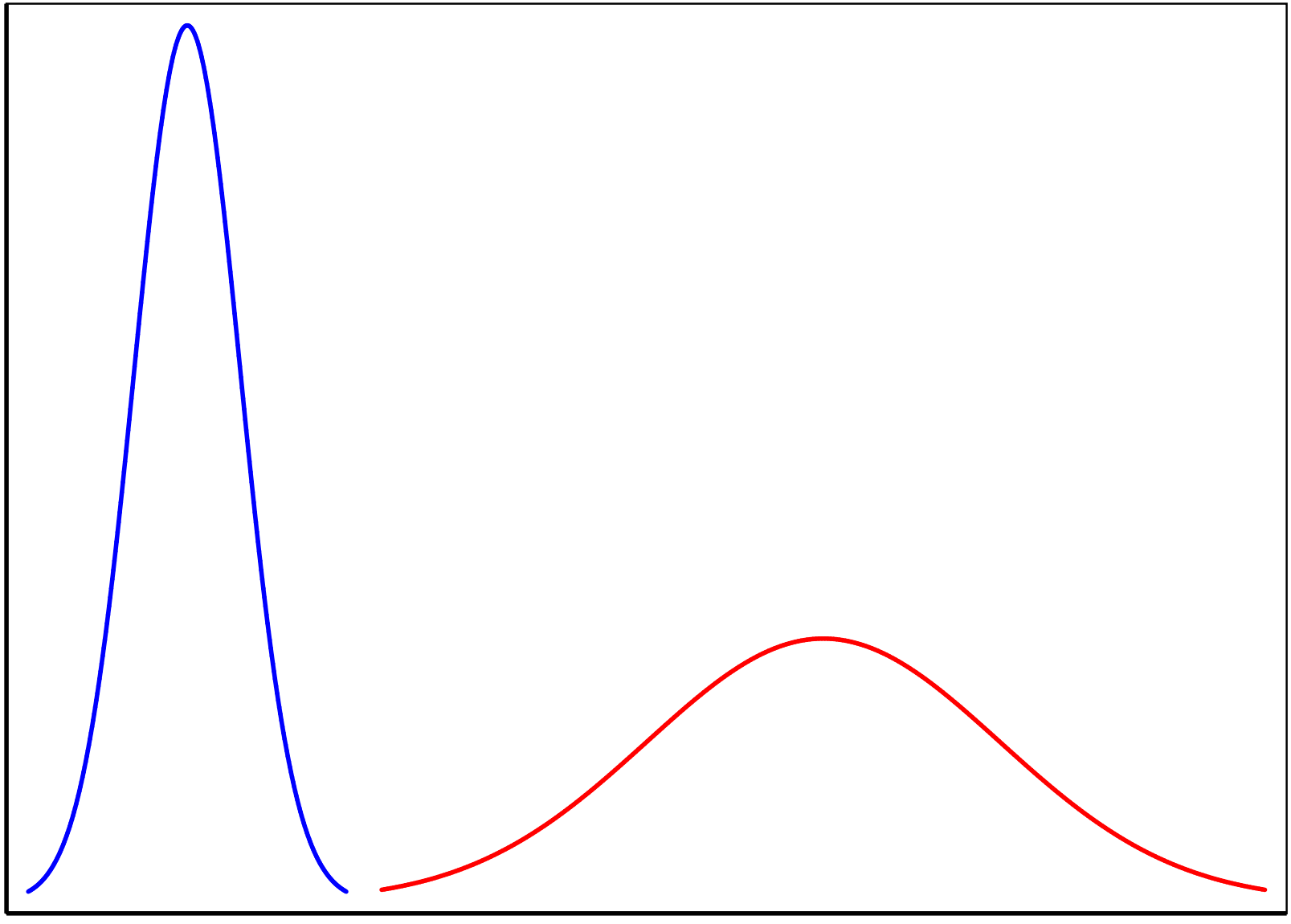
Sampling Distribution: Bigger or Smaller Standard Error?



Sampling Distribution: Bigger or Smaller Standard Error?



Sampling Distribution: Which One Do We Prefer?



What is Power?

What is Power?

- The ability of our experiment to detect statistically significant treatment effects, if they really exist
- Experiment's ability to avoid making a Type II error (incorrect failure to reject the null hypothesis of no effect). Pregnancy example?
- The probability of being in the rejection region of the null hypothesis if the alternative hypothesis is true

What is Power? Example

- John runs an experiment to see whether giving people cash makes them more likely to start a business compared to giving them loans
- Finds no statistically significant difference between the groups
- What does this mean?

Why Might an Under-Powered Study be Bad?

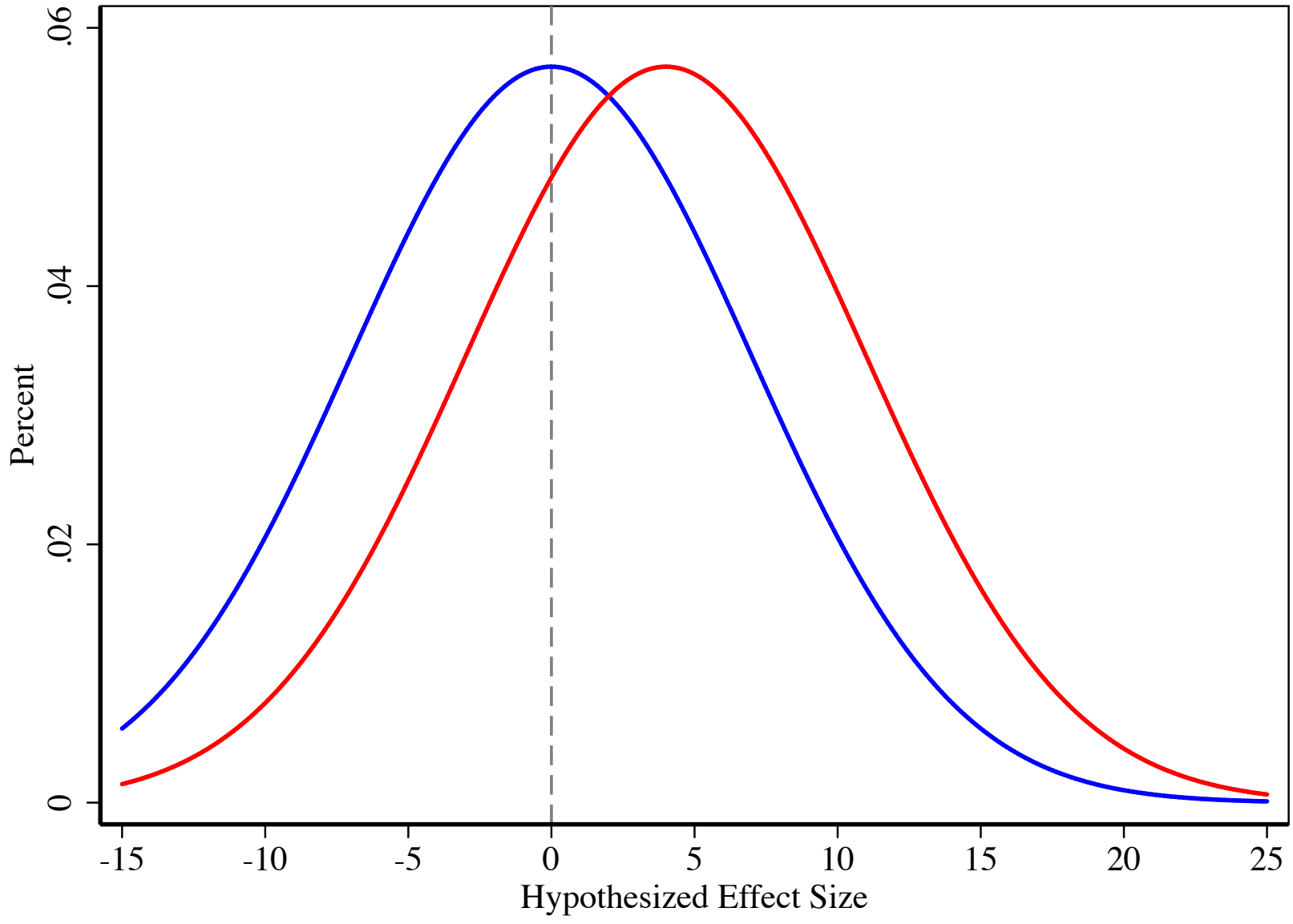
Why Might an Under-Powered Study be Bad?

Cost and interpretation

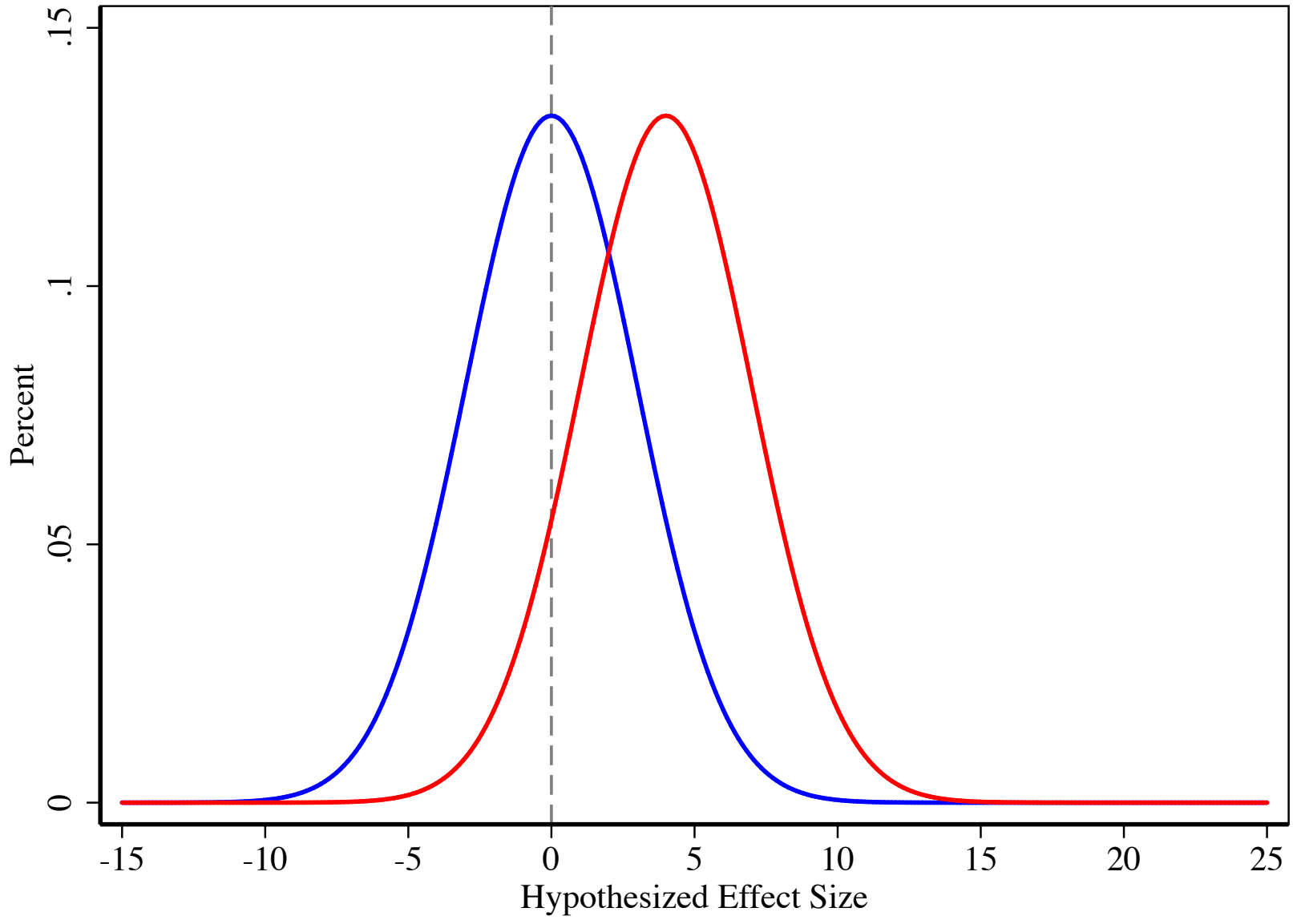
Starting Point for Power Analysis

- Power analysis is something we do before we run a study
- Goal: to discover whether our planned design has enough power to detect effects if they exist
- We usually state a hypothesis about the effect-size of a treatment and compare this against the null hypothesis of no effect
- Both the null and alternative hypotheses have associated sampling distributions which matter for power
- Let's see some examples. Which of the following are high-powered designs?

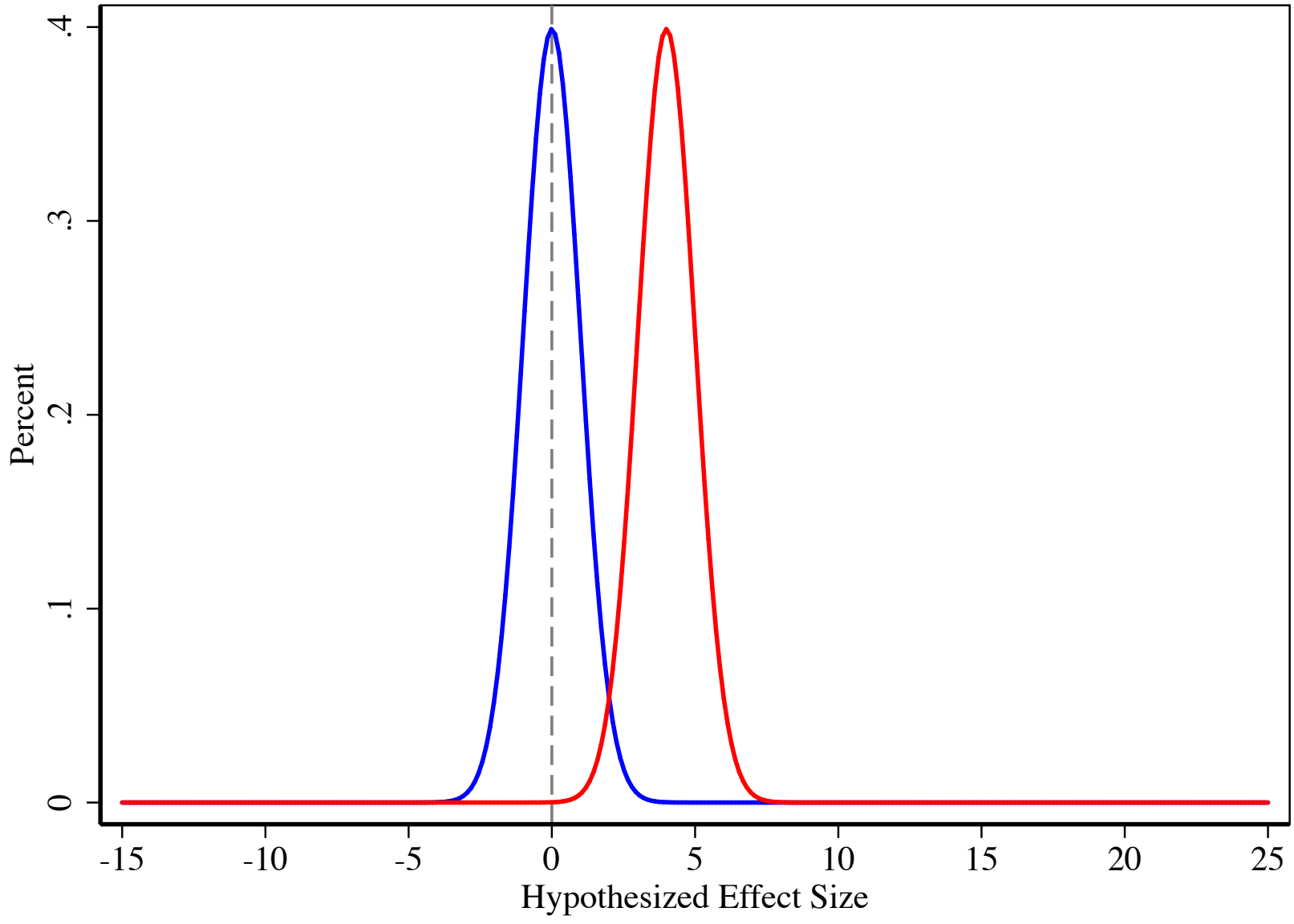
Graphical Intuition



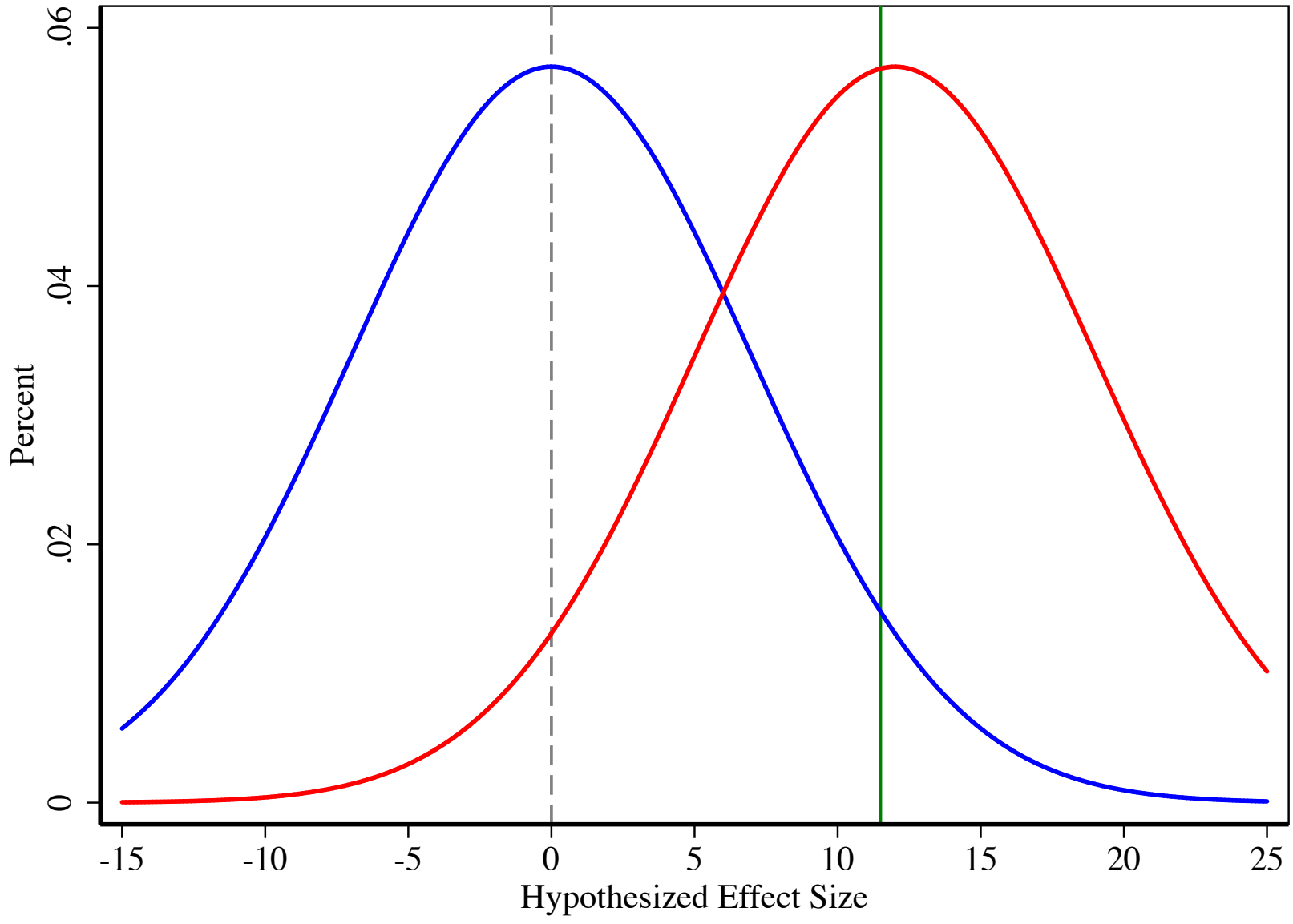
Graphical Intuition



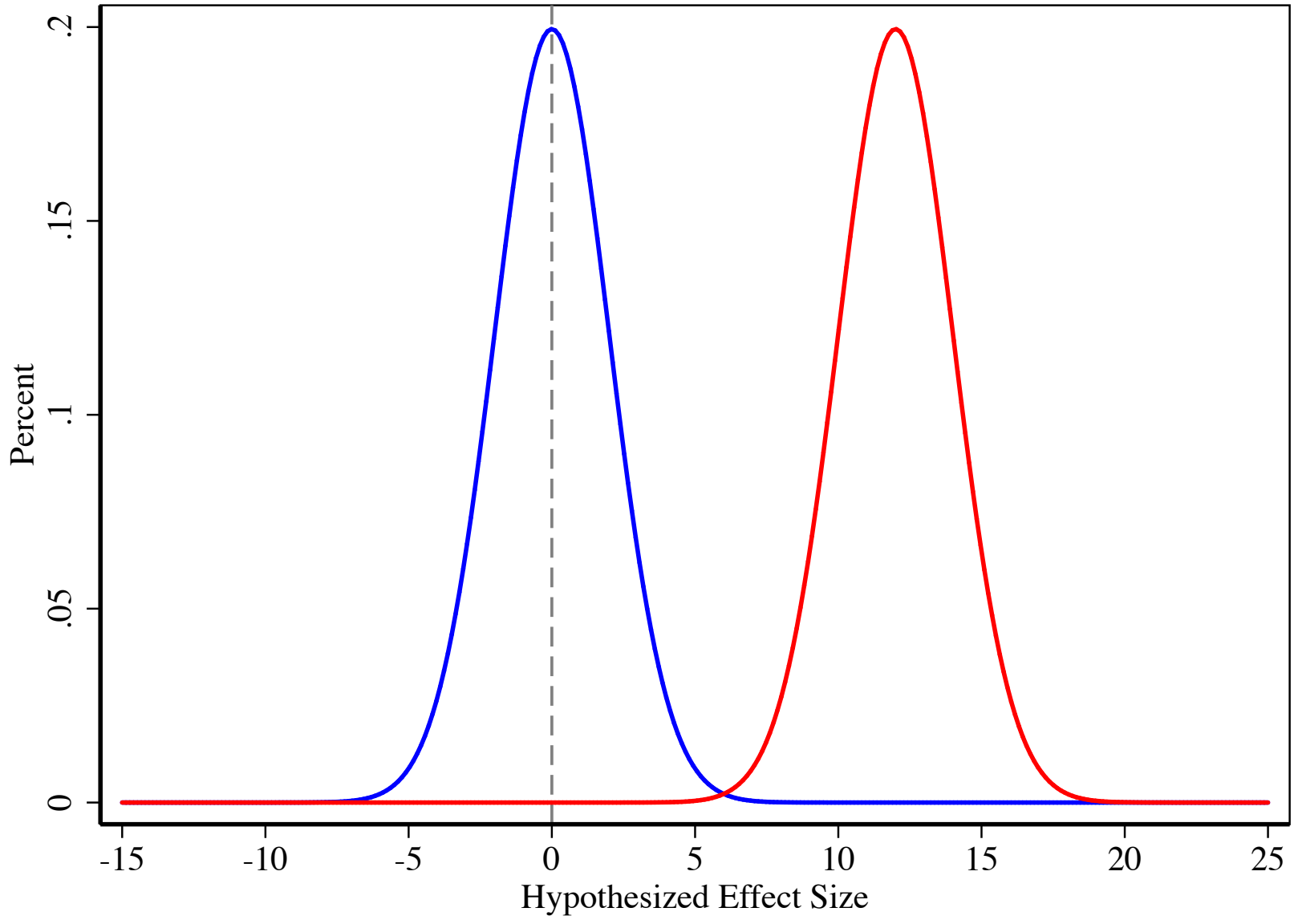
Graphical Intuition



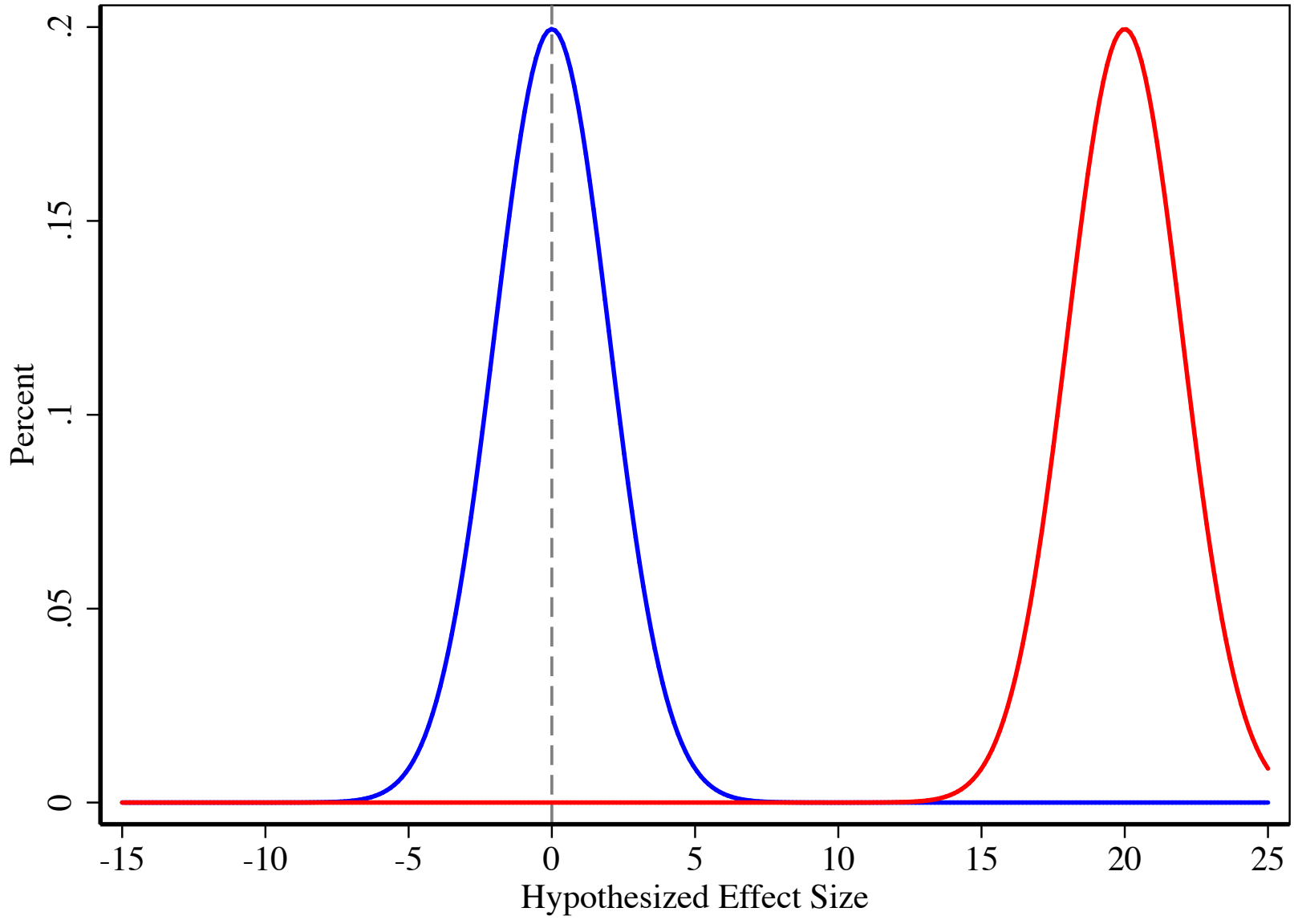
Graphical Intuition



Graphical Intuition



Graphical Intuition



What are the Three Main Inputs into Statistical Power?

What are the Three Main Inputs into Statistical Power?

- Sample size
- Noisiness of the outcome variable (σ)
- Treatment-effect size

The Power Formula

$$Power = \Phi\left(\frac{|\tau|\sqrt{N}}{2\sigma} - \Phi^{-1}(1 - \frac{\alpha}{2})\right) \quad (1)$$

- Power is a number between 0 and 1; higher is better
- Φ is the conditional density function of the normal distribution FIXED
- τ is the effect size
- N is the sample size
- σ is the standard deviation of the outcome
- α is the significance level FIXED (by convention)

Health warning: this makes many assumptions we haven't discussed so far

The Power Formula

$$Power = \Phi\left(\frac{|\tau|\sqrt{N}}{2\sigma} - \Phi^{-1}(1 - \frac{\alpha}{2})\right) \tag{2}$$

- Power is a number between 0 and 1; higher is better
- Φ is the conditional density function of the normal distribution FIXED
- τ is the effect size CAN CHANGE
- N is the sample size CAN CHANGE
- σ is the standard deviation of the outcome CAN CHANGE
- α is the significance level FIXED

Three Main Inputs into Statistical Power 1: Sample Size

- More observations $\rightarrow$ more power
- Add observations!
- Problems?

Three Main Inputs into Statistical Power 2: Noisiness of Outcome Measure

- Less noise → more power
- Reduce noise. How?
  - Blocking—conduct experiments among subjects that look more similar
  - Collect baseline covariates—background information about experimental units
  - Collect multiple measures of outcomes
- Problems?

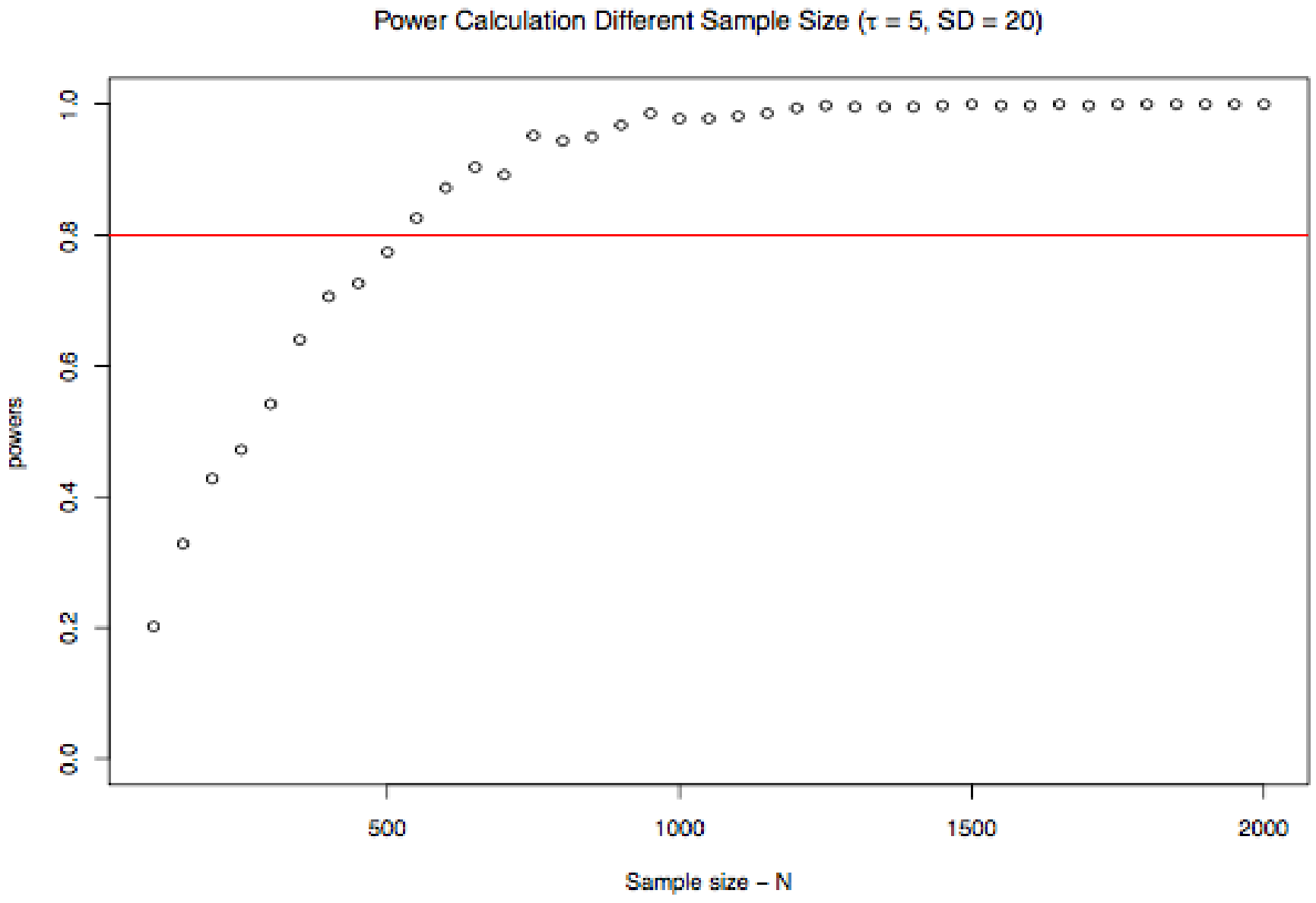
Three Main Inputs into Statistical Power 3: Size of Treatment Effect

- Bigger effect $\rightarrow$ more power
- Boost dosage / avoid very weak treatments
- Problems?

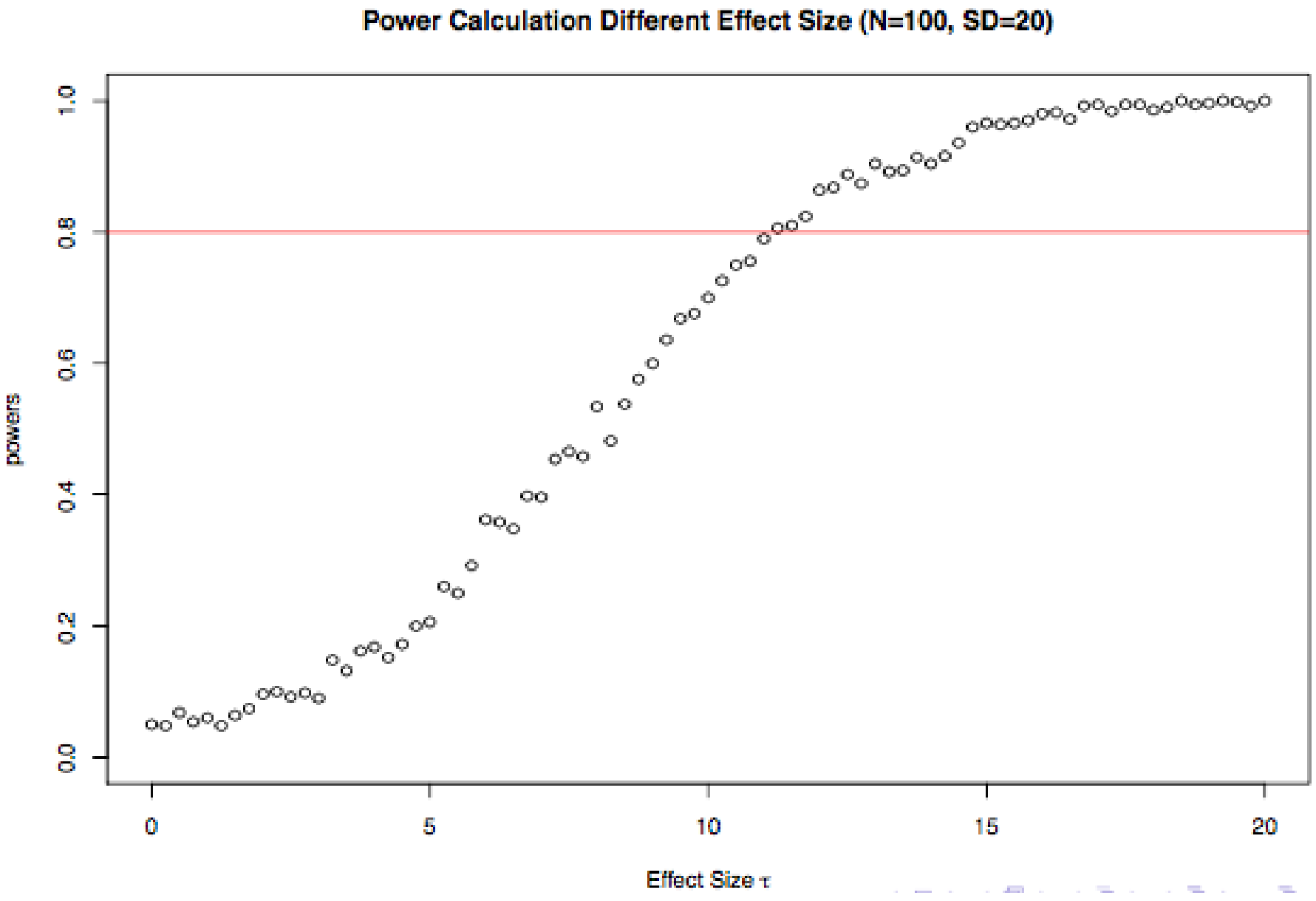
Power is the Art of Tweaking!

We tweak different parts of our design up front to make sure that our experiment has enough power to detect effects (assuming they exist)

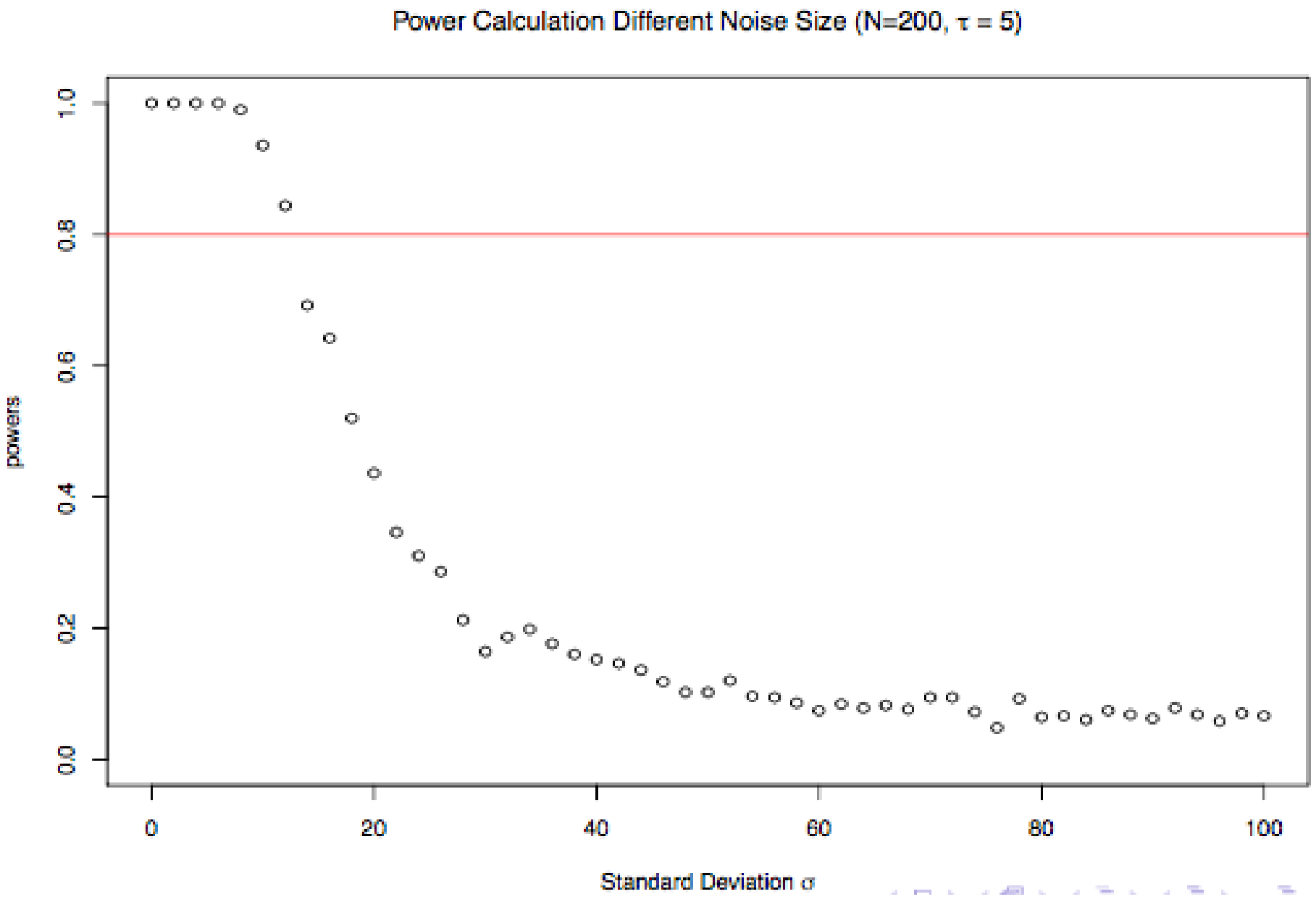
Tweak Sample Size: How Does Power Respond?



Tweak Effect Size: How Does Power Respond?



Tweak SD of Outcome: How Does Power Respond?



Your Turn!

- Go to `http://egap.org/`
- Tools $>$ Apps $>$ EGAP Tool: Power Calculator
- Set Significance Level at Alpha $= 0.05$
- Set Power Target at 0.8
- Set Maximum Number of Subjects at 1000

Your Turn!

Problems:

1. Fix Standard Deviation of Outcome Variable at 10. How many subjects do I need if my Treatment Effect Size is 2 in order for my experiment to have 80% power? What about Treatment Effect Size 5? Treatment Effect Size 10?
2. Fix Treatment Effect Size at 20. How many subjects do I need if the Standard Deviation of Outcome Variable is 10 in order for my experiment to have 80% power? What if the Standard Deviation of Outcome Variable is 20? 30? 100?

An Alternative Perspective: Minimum Detectable Effect

- Hardest part of power analysis is plugging in treatment effect—how can we possibly know before experiment has been run?
- Ask two questions:
  1. For a give set of inputs, what's the smallest effect that my study would be able to detect?
  2. Would this effect-size be "satisfactory"?

  - Cost-effectiveness
  - Disciplinary rules of thumb (e.g. 0.2 SD effects in education research)
  - Other studies which had similar goals to yours
- Remember: Small effects are harder to detect than big effects!

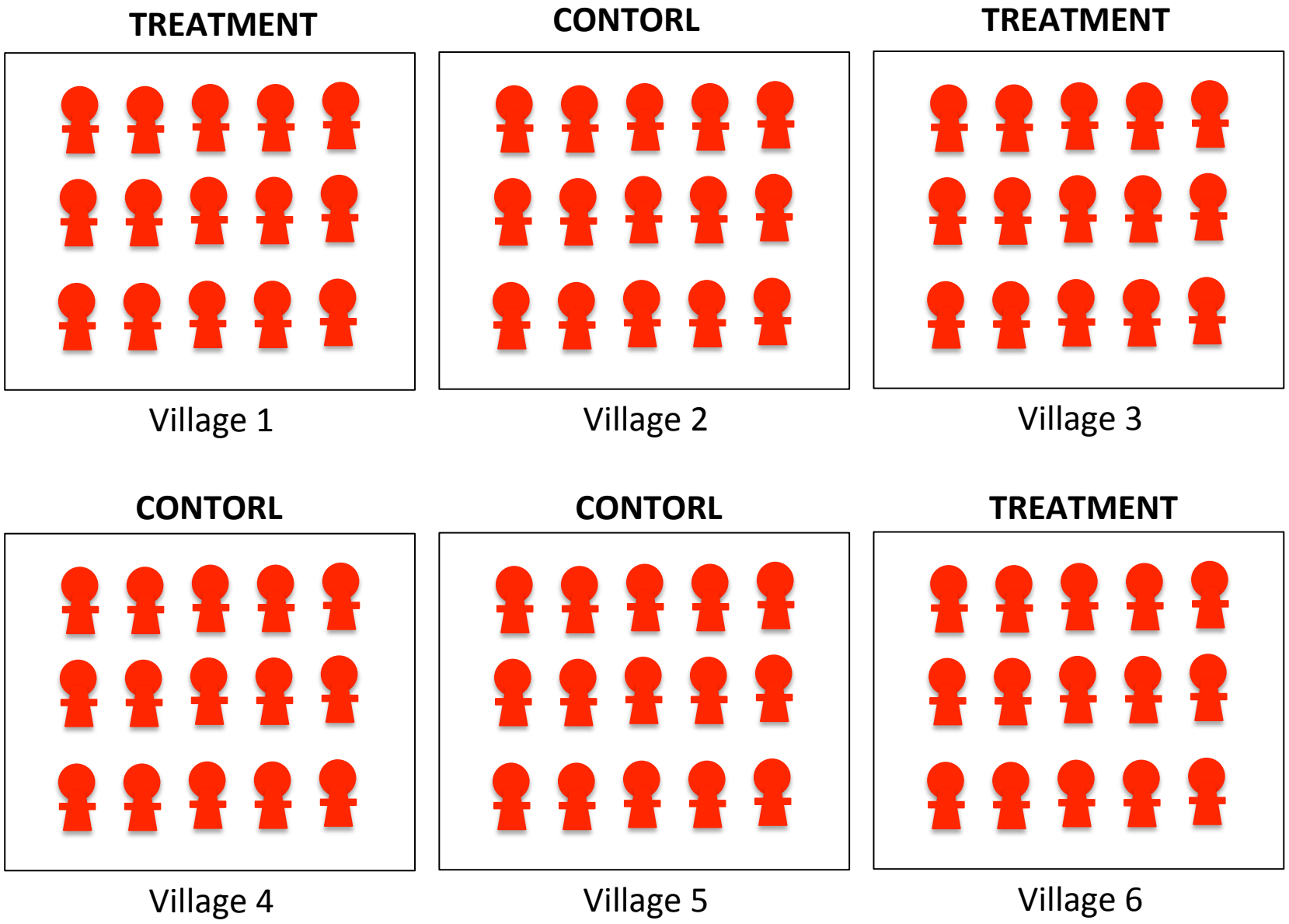
An Alternative Perspective: Minimum Detectable Effect

$$|MDE| = (t_{\alpha/2} + t_{1-\kappa})\sigma_{\hat{\beta}} \tag{3}$$

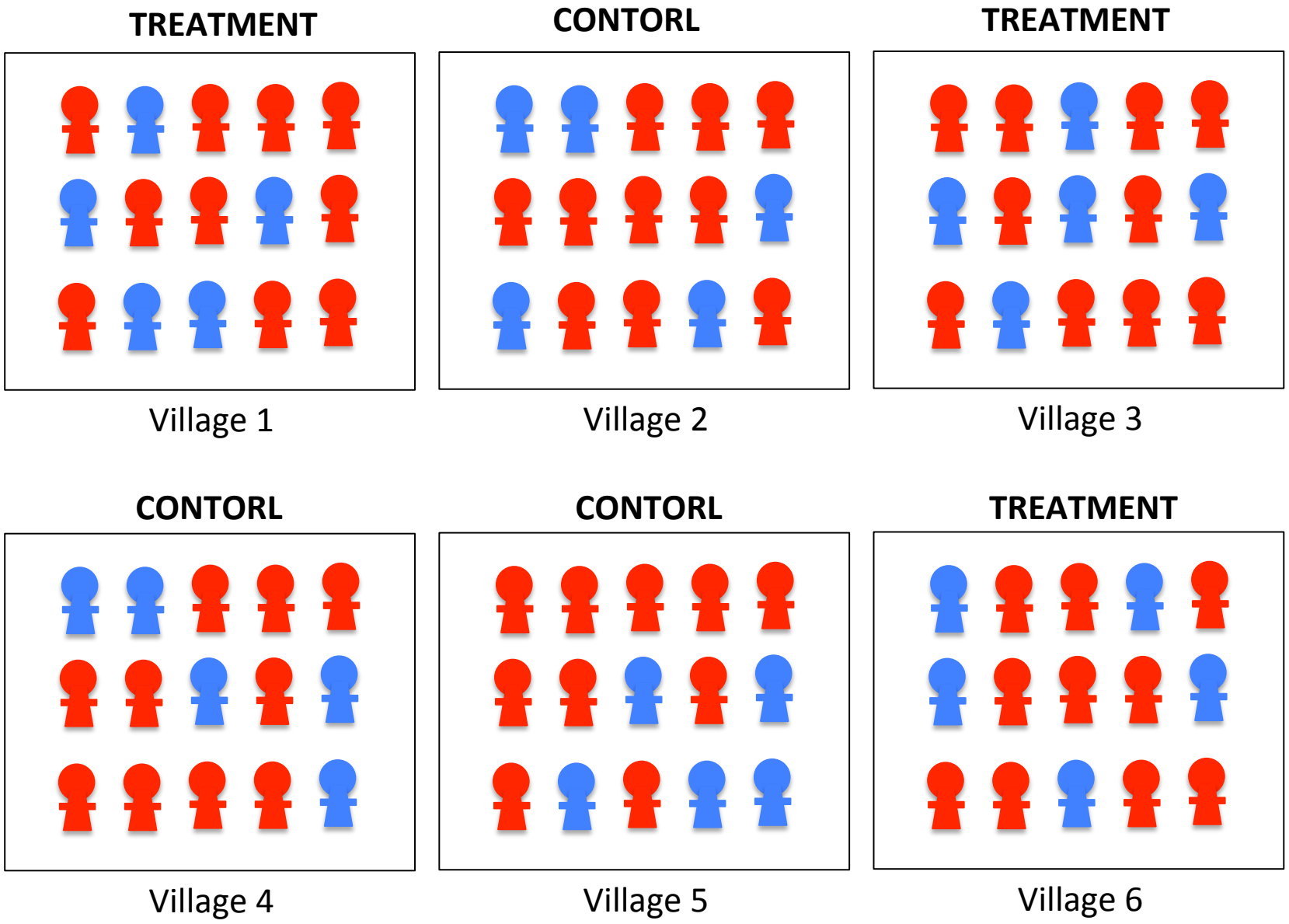
- Fix α at 0.05 and κ at 0.80 (industry standards)
- $t_{\alpha/2}$ and $t_{1-\kappa}$ are absolute values of relevant quantiles of the test statistic. Because most test statistics are normally distributed, $t_{\alpha/2} + t_{1-\kappa} = |z_{0.25}| + |z_{0.20}| = 1.96 + 0.84 = 2.80$

Special Case: Clustered-Randomized Designs

Special Case: Clustered-Randomized Designs



Special Case: Clustered-Randomized Designs



Special Case: Clustered-Randomized Designs

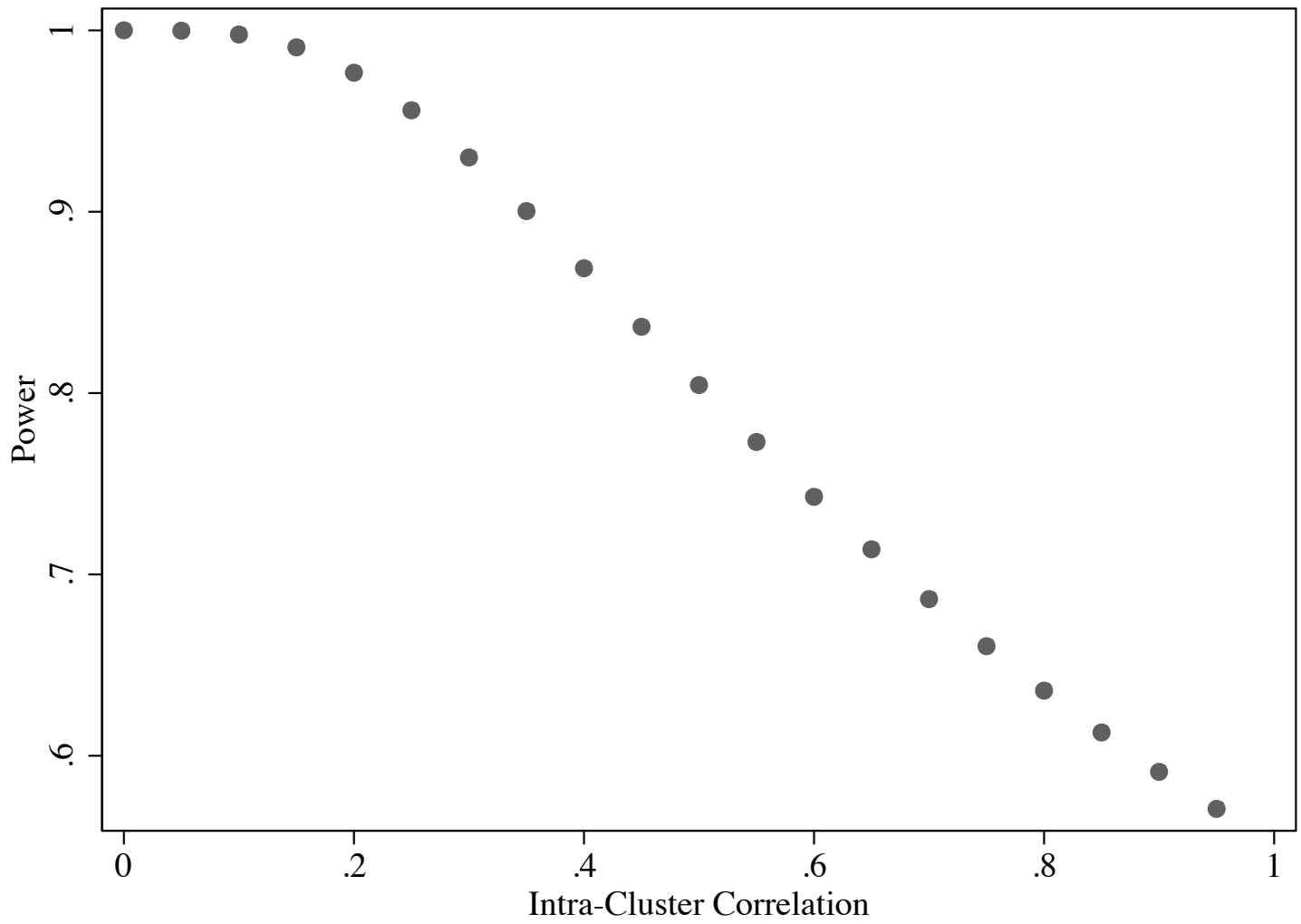
- Used if intervention has to function at the cluster level or if outcome defined at the cluster level
- We often want to randomize treatment at the level of groups, but only have the ability to sample a few people within those groups
- Examples?
- Special issues for power:
  - Number of individuals sampled per cluster
  - Intra-cluster correlation

Intra-Cluster Correlation: What is it?

- To what extent can we predict people's outcomes based on which group they're in? Is the clustering important for people's outcomes?
- Example:
  - 2000 students, divided into 100 classes of 20 students each; 1/2 classes in treatment, 1/2 control
  - When the intracluster correlation is 0, individuals within classes are no more similar than individuals in different classes
  - It's like assigning 2000 individuals to treatment or control!
  - When the intracluster correlation is 1, everyone within a class acts the same, and so you effectively have 100 independent observations
  - Implications for power?

Tweak Intra-Cluster Correlation: How Does Power Respond?

Number of clusters = 140; 10 sampled per cluster

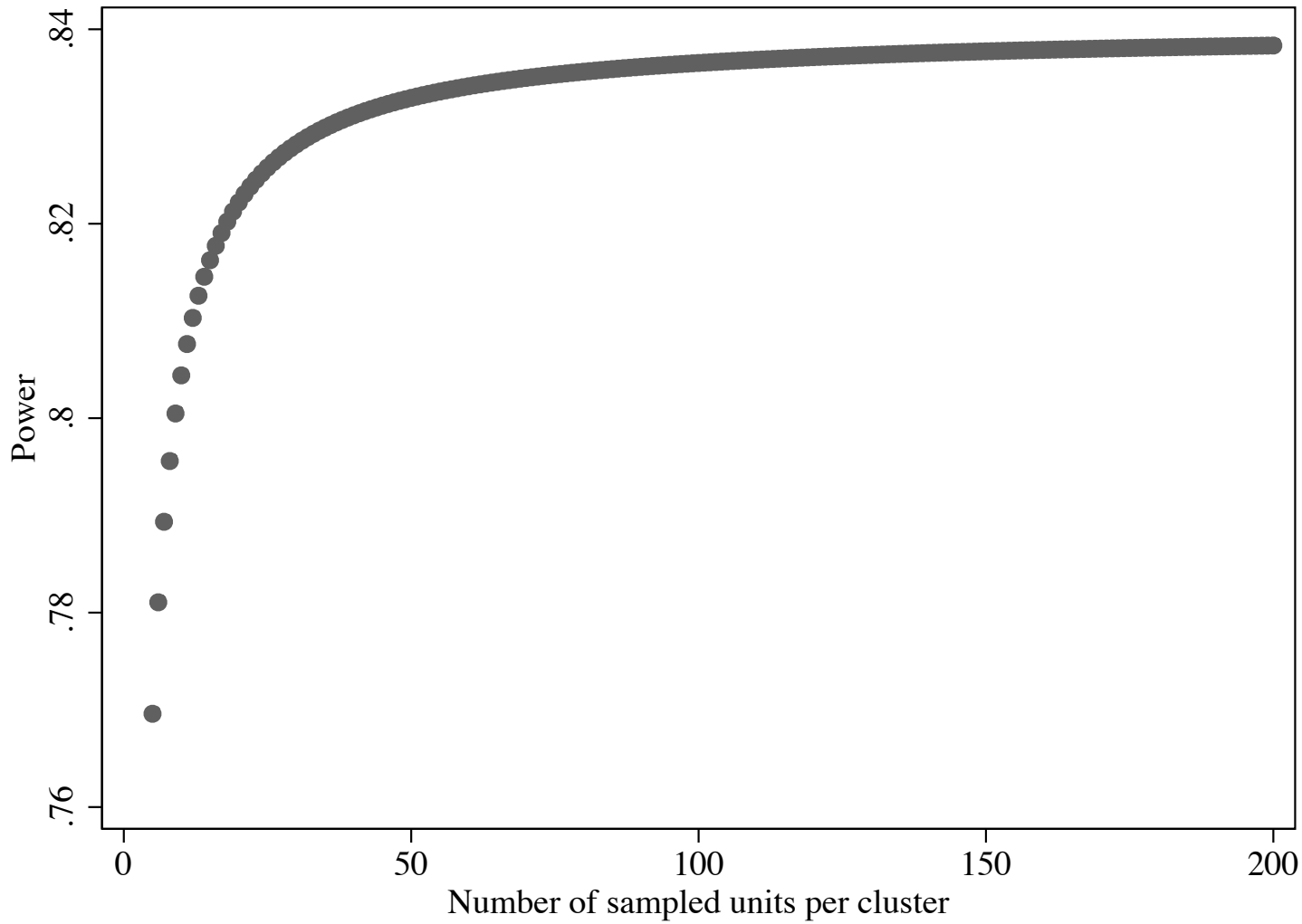


Tweak Number of Units Per Cluster: How Does Power Respond?

- Another choice we have to make in cluster designs is how many units *within* clusters to sample
- Surely we want to sample as many as possible, right?
- Hmm...

Tweak Number of Units Per Cluster: How Does Power Respond?

ICC = 0.5, number of clusters = 140



Golden Rule of Cluster-Randomized Designs

Unless intra-cluster correlation is very small, it's always better to add more clusters than to sample more people within the clusters

Your Turn!

- Go to `http://egap.org/`
- Tools > Apps > EGAP Tool: Power Calculator
- Click box which says "Clustered Design?"
- Set Significance Level at Alpha = 0.05
- Set Treatment Effect Size at 5
- Standard Deviation of Outcome Variable at 10
- Set Power Target at 0.8
- Set Maximum Number of Subjects at 2000

Your Turn!

Problems:

1. Fix Number of Clusters per Arm at 40. How many subjects do I need if my Intra-cluster Correlation is 0.6 in order for my experiment to have 80% power? What about Intra-cluster Correlation of 0.4? 0.1? 0?
2. Fix Intra-cluster Correlation at 0.5. How many subjects do I need if the Number of Clusters per Arm is 100 in order for my experiment to have 80% power? What is the Number of Clusters per Arm is 50? 35? 20?

Recap: What Have you Learned?

Takeaways

- Power is the ability of our experiment to detect statistically significant treatment effects, if they in fact exist
- Power matters: for practical reasons and for interpretation
- Increase power by strengthening intervention, reducing noise, and increasing sample size
- In cluster-randomized designs, almost always better to add more clusters rather than interview more people within clusters
- Always run a power analysis before committing to a final design
- But beware that it involves some guesswork; be skeptical and vary assumptions

References

Note, several of these slides are not original. Material is borrowed from several sources:

- Cyrus Samii, NYU slides (on minimum detectable effects)
- Tara Slough, Columbia slides (graphs on the sensitivity of effects)
- World Bank Impact Evaluation Blog (for description of ICC)
- Glennester book, especially the chapter on power