Short Course on Program Evaluation
Power Calculations and Sampling Design

Matias D. Cattaneo

University of Michigan

May 31, 2017

Outline

1. General Ideas
2. Experimental Design
3. Application: RD Designs

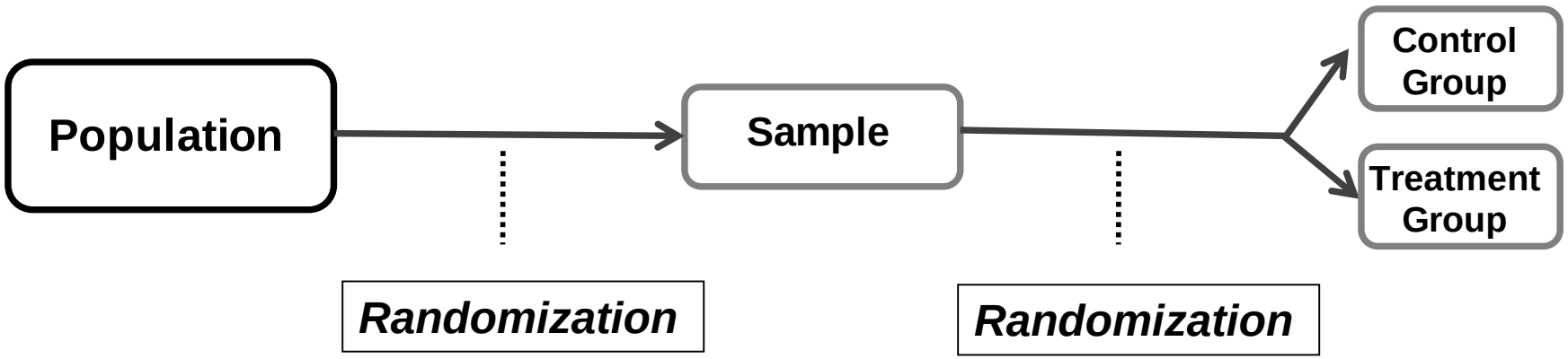
Randomization

- If the researcher assigns the subjects to the groups at random or by chance, the two groups will be on average balanced with respect to all observable and unobservable factors other than treatment.

- In principle, randomized trials ensure that outcomes in the control group really do capture the counterfactual for a treatment group.

- Random assignment is achieved by any procedure that assigns units to conditions based only on chance (toss of a coin, random numbers), in which each unit has the same nonzero probability of being assigned to a condition.

Two-Stage Randomization



1st Stage:

Ensures that the results in the sample will represent the results in the population within a defined level of sampling error

External Validity

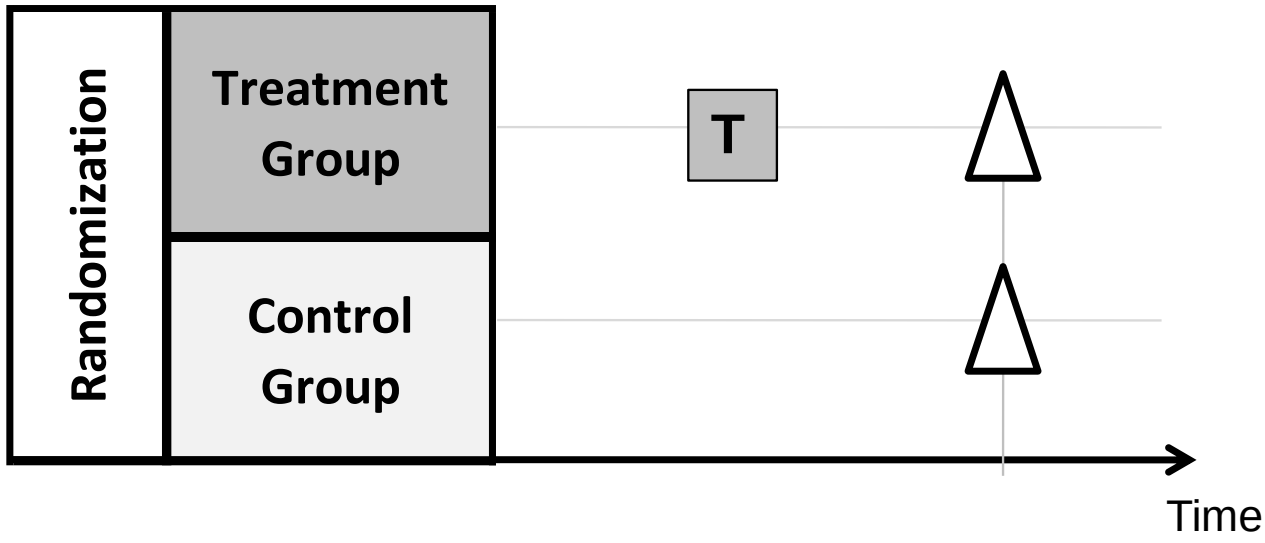
2nd Stage:

Ensures that the observed effect on the dependent variable is due to some aspect of the treatment rather than other confounding factors

Internal Validity

Designs with Random Assignment

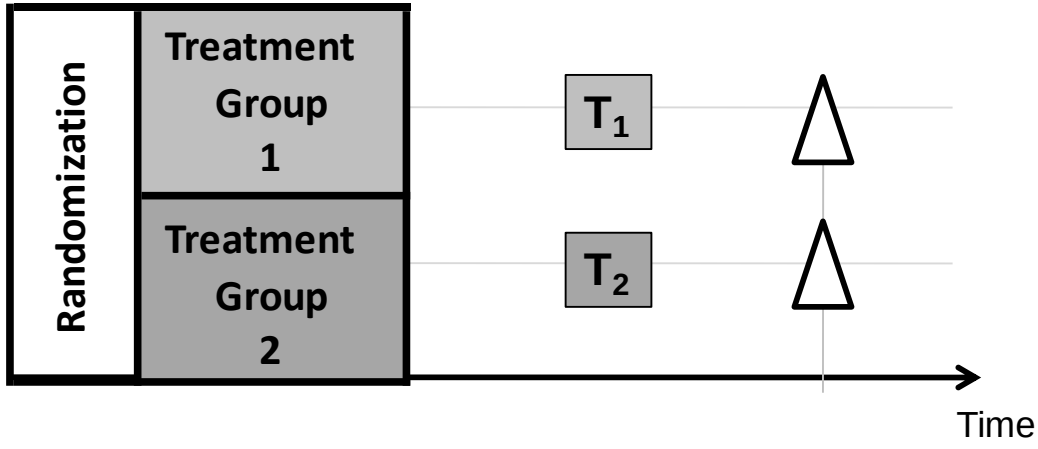
• **<u>Basic Design</u>:** random assignment of units to treatment and control groups.



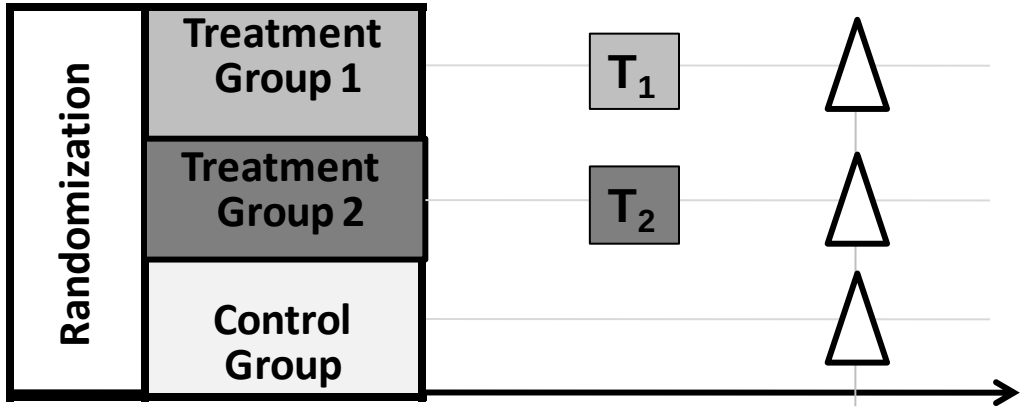
Designs with Random Assignment

- **<u>Variants to the Basic Design</u>:**

Two Alternative Treatments

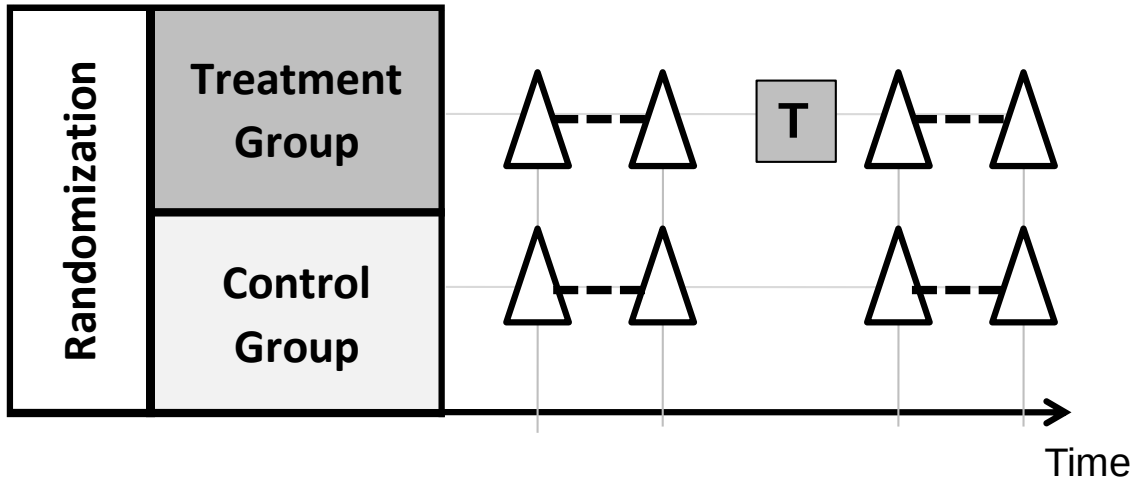


Two Alternative Treatments and Control



Designs with Random Assignment

- **Longitudinal Design:**



- In practice it is vary rare to have many pre-tests.

- Not always feasible to have many post-tests.

  - Increases risk of units leaving the experiment (attrition).
  - In control groups some units will receive treatment (either this or other one).
  - In some cases, not ethical to deprive units from a beneficial treatment.

Designs with Random Assignment

• **Factorial Design:** provides the possibility of identifying the effect of two or more independent variables (factors), each with two or more levels of intensity.

•This would be a 2x2 design. Units are assigned to one of these four cells:

		Factor B		
		Level 1	**Level 2**	
Factor A	**Level 1**	***A1B1***	***A1B2***	**A1**
	Level 2	***A2B1***	***A2B2***	**A2**
		B1	**B2**	

• Advantage of factorial design:

  • Allows testing combinations of treatments easily.

Power Analysis

- An important element in the design of an evaluation study is the determination of the sample sizes needed to reject confidently the null hypothesis of no program impact or, equivalently, to detect an impact of some minimal magnitude with a given level of confidence.

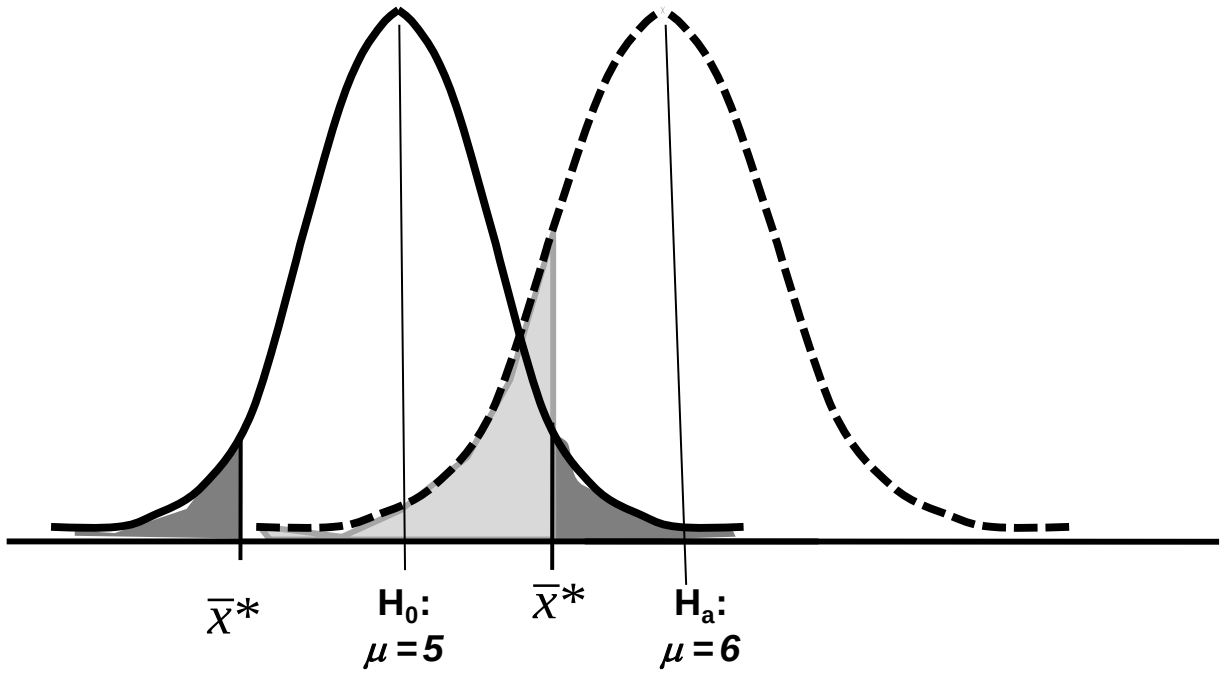
- There are two different types of errors we can incur when testing hypothesis: type I and type II.

- The probability α is called type-I-error rate, since it denotes the probability of rejecting the null when it is true.

- The second type of error we can make consists of retaining H_0, when Ha is true. This error is named β or type II error.

	H_0 Accepted	H_0 Rejected
H_0 true	✓	Type I error
H_1 true	Type II error	✓

Type I and Type II errors



Power of a Test

- Denote the probability of failing to reject the null when it is false as β; then, $1 - \beta$ is the probability of correctly rejecting the null (not incurring in type II error).

- This last probability is called the power of the test. The maximum power a test can have is 1, the minimum is 0. Ideally we want a test to have high power, close to 1.

- Note that the power of the test depends on the rule that we fix to accept or reject the null and on the true distribution of the statistic. However, the true distribution of the statistic is unknown since it depends on the value of the parameter m which is exactly what we are trying to estimate.

- Since we dont know the true m, we could not possibly know the true distribution of the statistic; and thus, we cannot calculate the power of the test.

- But what we can do, is estimate the power of the test under different alternative hypothesis.

- As we can deduce from the next figure, given a value for , as we set an alternative hypothesis that is far away from the null, the power of the test increases (i.e., the type II error rate falls).

Power of a Test

What other factors affect the power of the test, given H_0 and H_A

- The probability of making a Type I error. If you increase the significance level , the areas where the null hypothesis will be rejected are larger.
- This increases the probably of rejecting the null hypothesis, and in particular if H_A is true.
- This reduces β, and hence, it increases the power of the test.
- Given α, any component that generates an increase in the probability of rejecting the null by increasing the test statistic will reduce and increase the power or the test.
- Those components are:
  - the sample size (the bigger, the more powerful is the test because the lower is the variance of the estimator);
  - the variance of the variable of interest (the smaller the population variability, the more powerful is the test, again, because the lower is the variance of the estimator).

Power of a Test

Thus, the power of a test increases with:

1. The distance between the value of the null and alternative hypothesis.
2. The significance level of the test.
3. Sample size.
4. Population variability of the variable of interest (or the residual of the model).

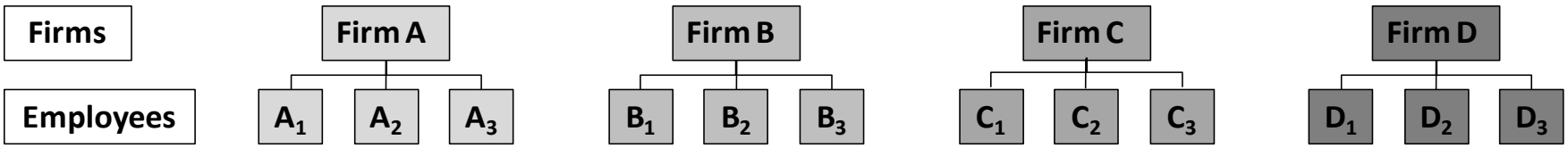
Experimental Design: Randomization Level

Randomization can be performed at different levels:

1. We can randomize the single units we observed.
   - Then, each unit has the same probability of being assigned to any treatment and we assumed that the observations were independent.
   - Randomization at unit-level: Simple Randomized Trials (SRT).
2. We can gather units into groups so that each unit belongs to one and only one group, and randomize over groups of units.
   - This is the case called Group Randomized Trials (GRT).
   - Randomization at group or cluster-level: Group Randomized Trials (GRT).

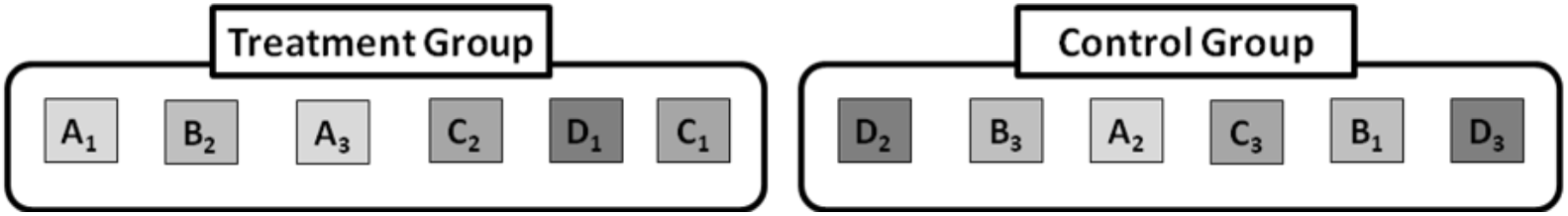
SRT versus GRT

- Consider an in-job training programs. The sample consists of 12 employees from 4 firms, which will be randomly assigned to treatment and control groups.

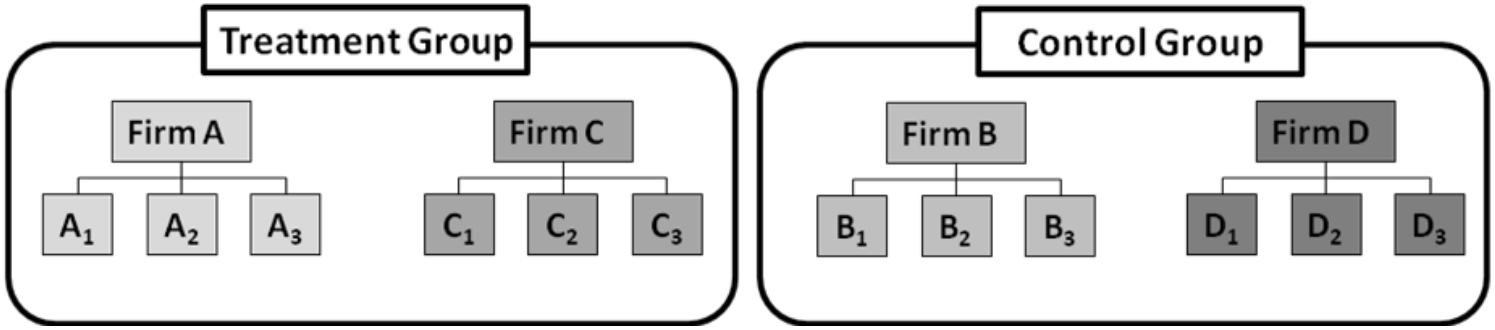


SRT versus GRT

- If units (employees) are randomly assigned to treatment – SRT-:



- If groups (firms) are randomly assigned to treatment – GRT-**:**



SRT versus GRT

- **There are three sources of variations in outcome measures:**

1. **Group effect:** all members belonging to a group share a common error term that captures any group characteristic unobservable to the researchers.

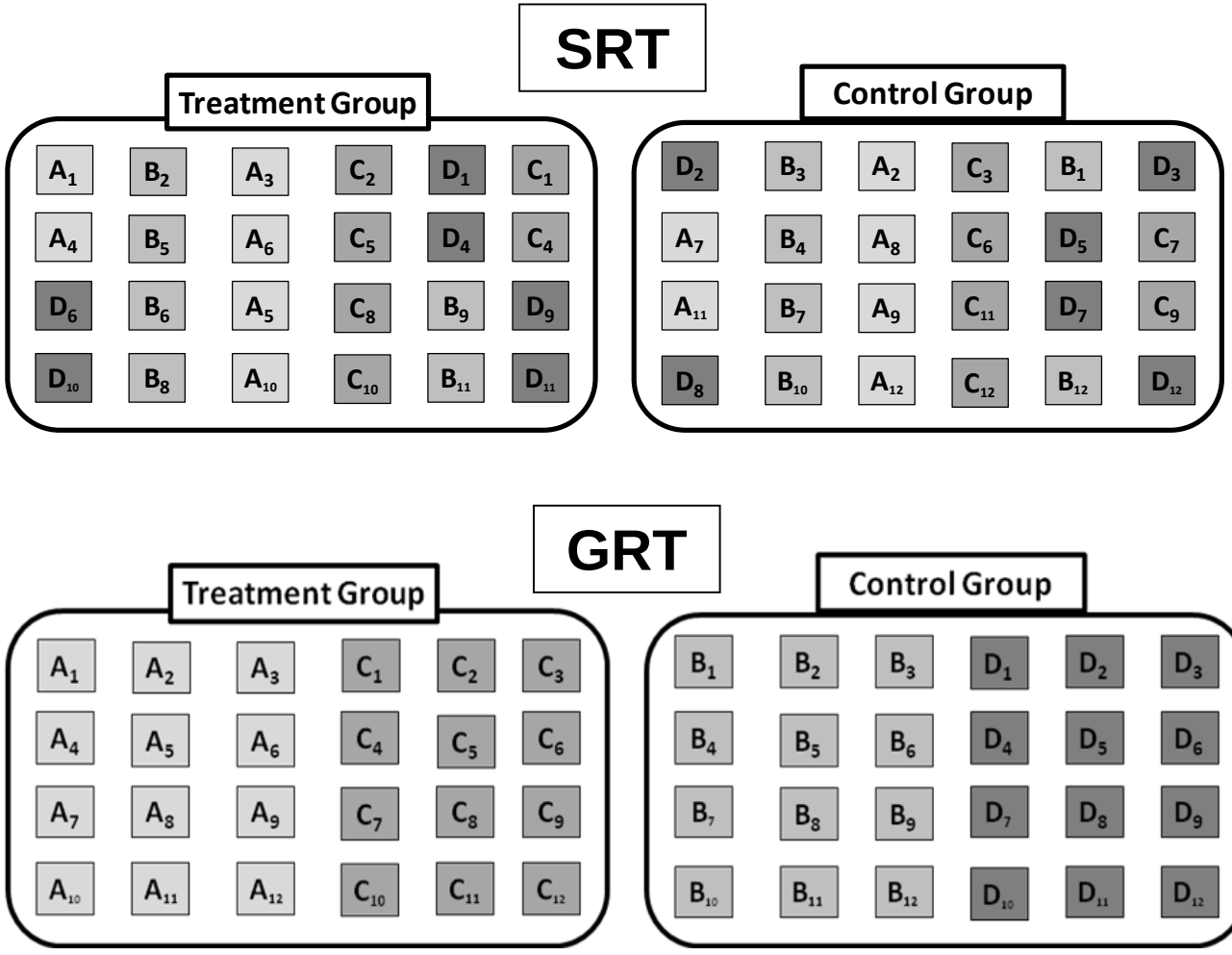
2. **Individual error term:** captures individual unobservable characteristics.

3. **Treatment effect:** due to different assignment to conditions.

Our aim is to isolate 3. We would like to reduce 1 and 2 to detect 3.

SRT versus GRT

- **Increasing the sample size, does not reduce the group term source of variation in outcomes.**



- Individual error terms cancel out in each group. And group errors are balanced across groups.

- Individual error terms cancel out in each group. But group errors are not balanced across groups, even if sample size increases.

Outline

1 General Ideas

2 Experimental Design

3 Application: RD Designs

Experimental Design: Relative Sample Sizes

- Suppose that you have n experimental subjects and you have to decide how many will be in the treatment group and how many in the control group.

- We know that:

$$\bar{Y}_1 - \bar{Y}_0 \sim \left(\mu_1 - \mu_0, \frac{\sigma_1^2}{N_1} + \frac{\sigma_0^2}{N_0}\right), \qquad n = N_0 + N_1.$$

- We want to choose N_0 and N_1 to minimize the variance of the estimator of the average treatment effect.

- The variance of $\bar{Y}_1 - \bar{Y}_0$ is

$$\mathbb{V}[\bar{Y}_1 - \bar{Y}_0] = \frac{\sigma_1^2}{pN} + \frac{\sigma_0^2}{(1-p)N}, \qquad p = \frac{N_1}{n},$$

  and therefore

$$p^* = \frac{\sigma_1}{\sigma_1 + \sigma_0} = \frac{1}{1 + \sigma_0/\sigma_1} \quad \Rightarrow \quad p^* \approx 0.5 \quad \text{if} \quad \sigma_0 \approx \sigma_1.$$

Experimental Design: Power Calculations to Choose Sample Size

- Recall that for a statistical test:
  - Type I error: Rejecting the null if the null is true.
  - Type II error: Not rejecting the null if the null is false.
- Size of a test is the probability of type I error, usually $\alpha = 0.05$.
- Power of a test is one minus the probability of type II error, i.e. the probability of rejecting the null if the null is false.
  - Statistical power increases with the sample size. But when is a sample "large enough"?
- We want to find n such that we will be able to detect an average treatment effect of size τ_1 or larger with high probability.
- Recall: under $H_0 : \tau_{\texttt{ATE}} = \tau_0$,

$$\frac{\hat{\tau} - \tau_0}{\text{s.e.}(\hat{\tau})} \sim \mathcal{N}(0,1), \qquad \hat{\tau} = \bar{Y}_1 - \bar{Y}_0, \qquad \text{s.e.}(\hat{\tau}) = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_0^2}{N_0}}$$

Experimental Design: Power Calculations to Choose Sample Size

- Power function, $H_0 : \tau_{\texttt{ATE}} = 0$, two-sided testing:

$$\beta(\tau) = \mathbb{P}_\tau[\text{Reject } H_0] = \mathbb{P}\left[\left|\frac{\hat{\tau} - \tau}{\text{s.e.}(\hat{\tau})}\right| > z_{1-\alpha/2}\right]$$

$$= \Phi\left(-z_{1-\alpha/2} - \frac{\tau}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_0^2}{N_0}}}\right) + 1 - \Phi\left(z_{1-\alpha/2} - \frac{\tau}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_0^2}{N_0}}}\right)$$

- Choose $n = N_1 + N_2$ we need to specify:
  - τ, minimum detectable magnitude of treatment effect.
  - $\beta(\tau)$, power value (usually 0.80 or higher).
  - σ_0^2 and σ_1^2 (usually use previous measurements, assume $\sigma_0^2 = \sigma_1^2$).
  - $p = N_1/n$, proportion of treatment units (if $\sigma_0^2 = \sigma_1^2$, then $p = 1/2$).

Outline

1 General Ideas

2 Experimental Design

3 Application: RD Designs

Power Calculations for RD

- Power Calculations with Misspecification Bias:

$$\hat{\beta}_\nu(\tau_{\mathtt{A}}) = 1 - \Phi\left(\frac{\tau_{\mathtt{A}} + \hat{\mathsf{B}}}{\sqrt{\hat{\mathsf{V}}}} + z_{1-\alpha/2}\right) + \Phi\left(\frac{\tau_{\mathtt{A}} + \hat{\mathsf{B}}}{\sqrt{\hat{\mathsf{V}}}} - z_{1-\alpha/2}\right),$$

where

$$\hat{\mathsf{B}} = \hat{h}_+^{1+p-\nu}\hat{\mathsf{B}}_+ - \hat{h}_-^{1+p-\nu}\hat{\mathsf{B}}_-, \qquad \hat{\mathsf{V}} = \frac{1}{n\hat{h}_+^{1+2\nu}}\hat{\mathsf{V}}_+ + \frac{1}{n\hat{h}_-^{1+2\nu}}\hat{\mathsf{V}}_-,$$

Note: if $\tau_{\mathtt{A}} = \tau_0$, then $\hat{\beta}_\nu(\tau_0) > \alpha$.

- Power Calculations with Robust Bias-Correction:

$$\hat{\beta}_\nu^{\mathsf{bc}}(\theta) = 1 - \Phi\left(\frac{\theta}{\sqrt{\hat{\mathsf{V}}^{\mathsf{bc}}}} + z_{1-\alpha/2}\right) + \Phi\left(\frac{\theta}{\sqrt{\hat{\mathsf{V}}^{\mathsf{bc}}}} - z_{1-\alpha/2}\right).$$

where

$$\hat{\theta}_\nu^{\mathsf{bc}} = \hat{\theta}_\nu - \hat{\mathsf{B}}, \qquad \hat{\mathsf{V}}^{\mathsf{bc}} = \frac{1}{n\hat{h}_+^{1+2\nu}}\hat{\mathsf{V}}_+^{\mathsf{bc}} + \frac{1}{n\hat{h}_-^{1+2\nu}}\hat{\mathsf{V}}_-^{\mathsf{bc}},$$

Effect of Changing Effective Sample Sizes

- Adjusted Power Function:

$$\tilde{\beta}_\nu^{\mathsf{bc}}(\theta) = 1 - \Phi\left(\frac{\theta}{\sqrt{\tilde{\mathsf{V}}^{\mathsf{bc}}}} + z_{1-\alpha/2}\right) + \Phi\left(\frac{\theta}{\sqrt{\tilde{\mathsf{V}}^{\mathsf{bc}}}} - z_{1-\alpha/2}\right),$$

$$\tilde{\mathsf{V}}^{\mathsf{bc}} = \frac{\hat{\mathsf{V}}_+^{\mathsf{bc}}}{m h_+^{1+2\nu}} + \frac{\hat{\mathsf{V}}_-^{\mathsf{bc}}}{m h_-^{1+2\nu}}, \qquad m = \frac{N_+}{N_{h_+}} \cdot M_+ + \frac{N_-}{N_{h_-}} \cdot M_-,$$

$$N_- = \sum_{i=1}^{n} \mathbb{1}(X_i < c), \quad N_+ = \sum_{i=1}^{n} \mathbb{1}(c \leq X_i),$$

$$N_{h_-} = \sum_{i=1}^{n} \mathbb{1}(c - h_- \leq X_i < c), \quad N_{h_+} = \sum_{i=1}^{n} \mathbb{1}(c \leq X_i \leq c + h_+),$$

  - M_- and M_+ denote the new (postulated by the user) sample sizes in the neighborhoods $[c - h_-, c)$ and $[c, c + h_+]$.
  - h_- and h_+ denote the new bandwidths chosen below and above the cutoff.
  - If $M_- = N_{h_-}$ and $M_+ = N_{h_+}$, then $m = n$.

Selecting New Sample Size

- Proportion of Treatment and Control Units:

$$\hat{\varrho}_\nu = \frac{\sqrt{\hat{\mathsf{V}}_+^{\mathsf{bc}}}}{\sqrt{\hat{\mathsf{V}}_-^{\mathsf{bc}}} + \sqrt{\hat{\mathsf{V}}_+^{\mathsf{bc}}}}.$$

  where

  - $M_1 = \hat{\varrho}_\nu M$ treatment units (i.e., those with $X_i \geq c$).
  - $M_0 = (1 - \hat{\varrho}_\nu) M$ control units (i.e., those with $X_i < c$).

- Selectionof Effective Observations $M = M_0 + M_1$:

$$M \text{ solves} \quad : \quad \beta = 1 - \Phi\left(\frac{\theta_{\mathsf{A}}}{\sqrt{\breve{\mathsf{V}}^{\mathsf{bc}}}} + z_{1-\alpha/2}\right) + \Phi\left(\frac{\theta_{\mathsf{A}}}{\sqrt{\breve{\mathsf{V}}^{\mathsf{bc}}}} - z_{1-\alpha/2}\right),$$

  where

$$\breve{\mathsf{V}}^{\mathsf{bc}} = \frac{1}{M} \cdot \frac{1}{\frac{N_+}{N_{h_+}} \cdot \hat{\varrho}_\nu + \frac{N_-}{N_{h_-}} \cdot (1 - \hat{\varrho}_\nu)} \cdot \left(\frac{\hat{\mathsf{V}}_+^{\mathsf{bc}}}{h_+^{1+2\nu}} + \frac{\hat{\mathsf{V}}_-^{\mathsf{bc}}}{h_-^{1+2\nu}}\right).$$