Power, Blocking, and Stratification

Mauricio Romero

Power, Blocking, and Stratification

Introduction

Statistical power

Blocking/Stratification

Relationship between Research Design and Analysis

Power, Blocking, and Stratification

Introduction

Statistical power

Blocking/Stratification

Relationship between Research Design and Analysis

Introduction

- In a simple experiment the average treatment effect is the difference in sample means between the treatment and the control group

- This is the OLS coefficient of β in the regression

$$Y_i = \alpha + \beta T_i + \varepsilon_i$$

Regression analysis of OLS

$$X'X = pN \begin{pmatrix} \frac{1}{p} & 1 \\ 1 & 1 \end{pmatrix}$$

$$\left(X'X\right)^{-1} = \frac{1}{N(1-p)} \begin{pmatrix} 1 & -1 \\ -1 & \frac{1}{p} \end{pmatrix}$$

And

$$V \begin{pmatrix} \widehat{\alpha} \\ \widehat{\delta} \end{pmatrix} = \sigma^2 (X'X)^{-1}$$

Power, Blocking, and Stratification

Introduction

Statistical power

Blocking/Stratification

Relationship between Research Design and Analysis

Power, Blocking, and Stratification

Introduction

Statistical power

Blocking/Stratification

Relationship between Research Design and Analysis

Statistical power

How many observations are enough?

Statistical power

How many observations are enough?

Definition

The power of the design is the probability that, for a given effect size and a given statistical significance level, we will be able to reject the hypothesis of zero effect

Statistical power

- **Is the unit of treatment the same as the unit of analysis?** Or, is the treatment to be administered to a ‘cluster’ of units?

Statistical power

- **Is the unit of treatment the same as the unit of analysis?** Or, is the treatment to be administered to a ‘cluster’ of units?
- Examples of individual randomizations:
  - Individuals who are given mobile phones to induce them to use an m-banking platform
  - Farmers individually provided with improved agricultural inputs
  - Students admitted to an elite school by a lottery process

Power, Blocking, and Stratification

Introduction

Statistical power

- Randomizing at the Unit of Analysis
- Cluster Randomized Experiments

Blocking/Stratification

Relationship between Research Design and Analysis

Randomizing at the Unit of Analysis

- The estimate of treatment effect is $\widehat{\beta}$ in the regression

$$Y_i = \alpha + \beta T_i + \varepsilon_i$$

- The mean of $\widehat{\beta}$ is β (the true effect)
- The variance of $\widehat{\beta}$ is $V(\widehat{\beta}) = \frac{\sigma^2}{p(1-p)N}$
- σ^2 is the variance of the outcome (Y_i)
- p is the proportion of treated units
- N is the number of observations

Randomizing at the Unit of Analysis

- We are generally interested in testing the null hypothesis (H_0) that the effect of the program is equal to zero against the alternative that it is not

- The **significance level**, or size, of a test represents the probability of a type I error, i.e., the probability we reject the hypothesis when it is in fact true

- The **power of the test** the probability that we reject H_0 when it is in fact false

Randomizing at the Unit of Analysis

- We are generally interested in testing the null hypothesis (H_0) that the effect of the program is equal to zero against the alternative that it is not
- The **significance level**, or size, of a test represents the probability of a type I error, i.e., the probability we reject the hypothesis when it is in fact true
- The **power of the test** the probability that we reject H_0 when it is in fact false

We will constantly use the fact that:

$$\widehat{\beta} \sim N\left(\beta, \frac{\sigma^2}{p(1-p)N}\right)$$

Randomizing at the Unit of Analysis

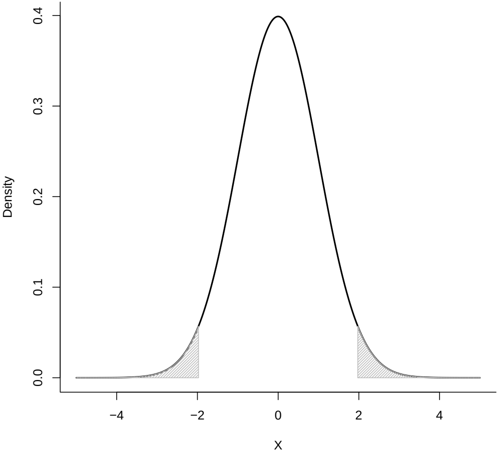
- We are generally interested in testing the null hypothesis (H_0) that the effect of the program is equal to zero against the alternative that it is not
- The **significance level**, or size, of a test represents the probability of a type I error, i.e., the probability we reject the hypothesis when it is in fact true
- The **power of the test** the probability that we reject H_0 when it is in fact false

We will constantly use the fact that:

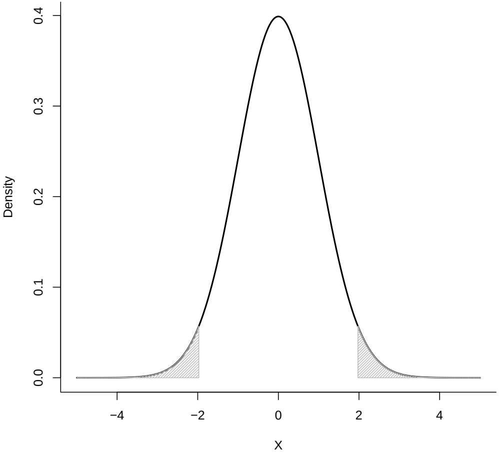
$$\widehat{\beta} \sim N\left(\beta, \frac{\sigma^2}{p(1-p)N}\right)$$

We often normalize the outcome and present results in terms of SD (so $\sigma^2 = 1$).

Significance level - Assume null is true (no effect)



Significance level - Assume null is true (no effect)

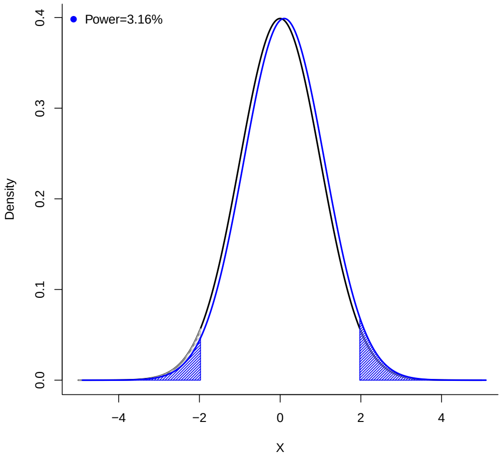


Gray area is the probability we reject the null when it is true

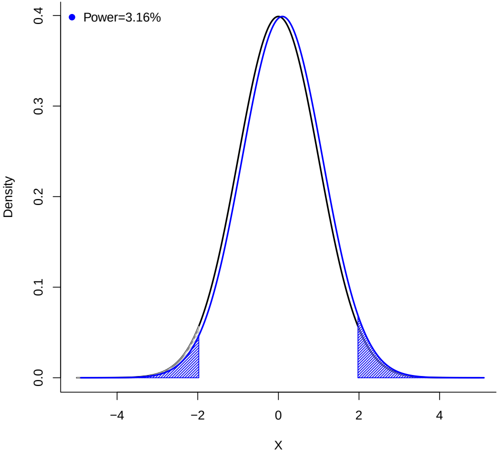
Power when the effect is β_1

For a true effect size β this is the fraction of the area under this curve that falls to the right of the critical value $t_{\frac{\alpha}{2}}$

Power when the effect is $\beta = 0.1$

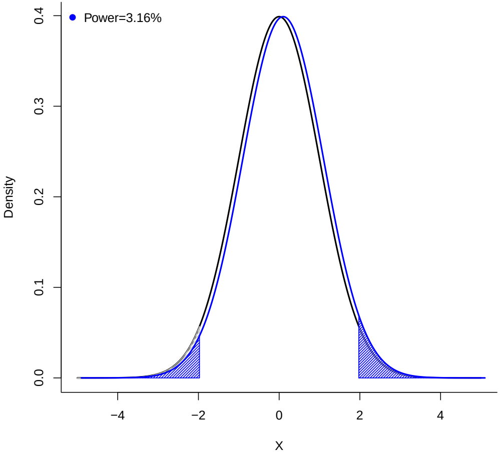


Power when the effect is $\beta = 0.1$

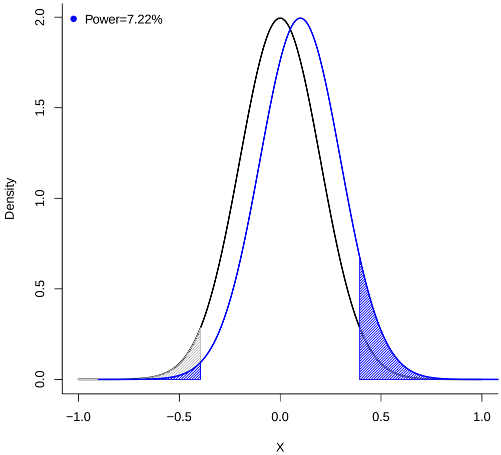


Blue area is the probability we reject the null when β is 0.1

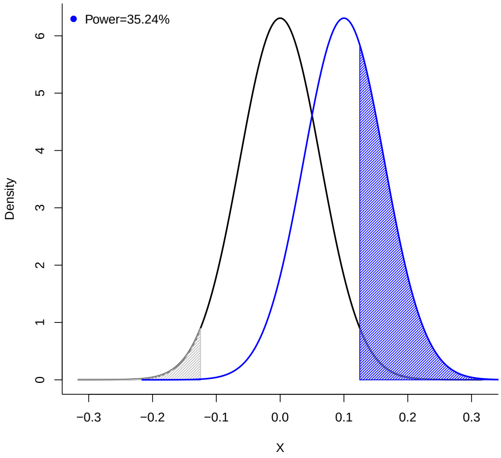
Power when $\beta_1 = 0.1$, $N = 4$, and $p = 0.5$



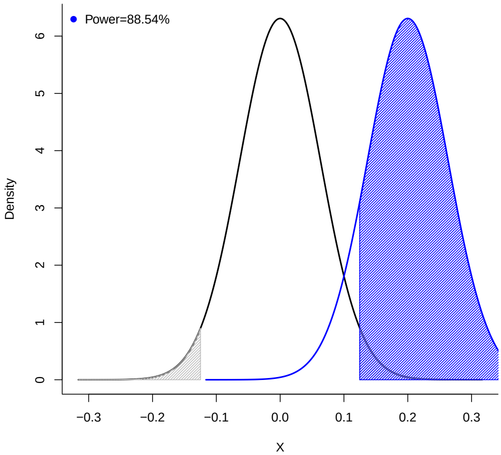
Power when $\beta_1 = 0.1$, $N = 100$, and $p = 0.5$



Power when $\beta_1 = 0.1$, $N = 1,000$, and $p = 0.5$

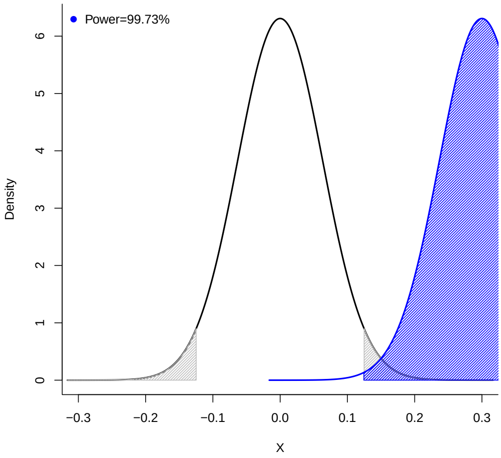


Power when $\beta = 0.2$, $N = 1,000$, and $p = 0.5$



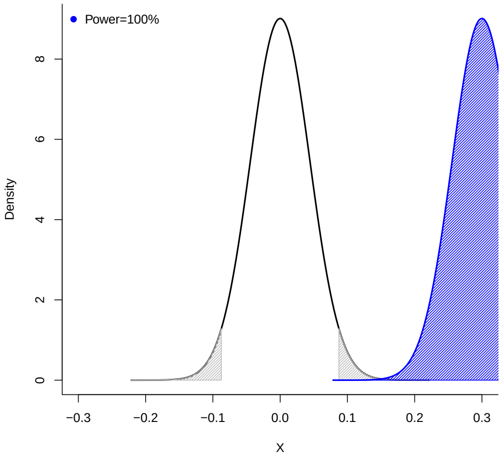
Blue area is the probability we reject the null when β is 0.2

Power when the effect is $\beta = 0.3$, $N = 1,000$, and $p = 0.5$



Blue area is the probability we reject the null when β is 0.3

Power when the effect is $\beta = 0.3$, $N = 1,000$, $p = 0.5$, and $\sigma = 0.7$



Blue area is the probability we reject the null when β is 0.3

Statistical power and clusters

- All these quantities we just looked at are related

- To achieve a power κ, it must therefore be that

$$\beta > (t_{\frac{\alpha}{2}} + t_{1-\kappa})\sigma_{\widehat{\beta}}$$

Minimum detectable effect

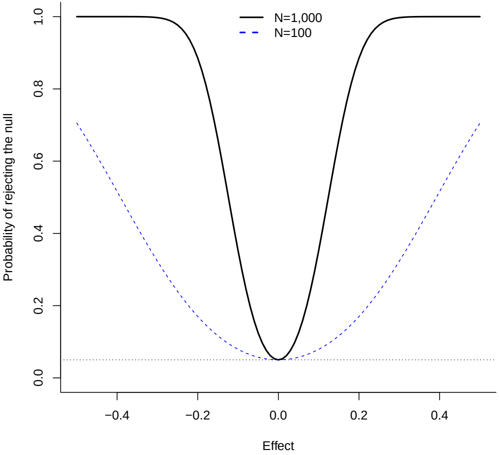
- The **minimum detectable effect** size for a given power (κ), significance level (α), sample size (N), and portion of subjects allocated to treatment group (p) is given by

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

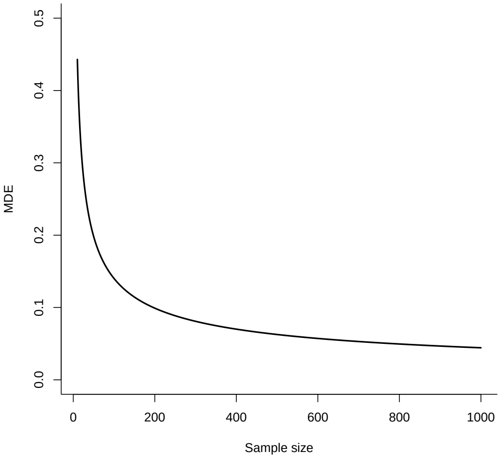
Randomizing at the Unit of Analysis

- The standard is to set $\kappa = 0.8$ or $\kappa = 0.9$
- The standard is to set $\alpha = 0.05$ or $\alpha = 0.1$
- The variance of outcomes σ^2 is typically the raw variance of the dependent variable you intend to use
- The sample size N is the number of observations in the study (you can change this)
- The fraction of the sample treated is p (you can change this)

Effect vs Power



Sample size vs MDE



How should you think about the MDE?

- What is the treatment effect below which it is pointless to implement the program?
- What is the minimum treatment effect that would make you willing to scale the program?
- If sample size is too small, you're likely to end up with an insignificant result for something that actually matters
- Small organizations often do not have the numbers to make an RCT worth conducting.

Power, Blocking, and Stratification

Introduction

Statistical power

- Randomizing at the Unit of Analysis
- Cluster Randomized Experiments

Blocking/Stratification

Relationship between Research Design and Analysis

Cluster Randomized Experiments

- Is the unit of treatment the same as the unit of analysis? Or, **is the treatment to be administered to a ‘cluster’ of units?**

Cluster Randomized Experiments

- Is the unit of treatment the same as the unit of analysis? Or, **is the treatment to be administered to a ‘cluster’ of units?**
- Examples of clustered randomizations:
  - Changing the business practices at a firm level and studying the impact on individual employees
  - Providing schools with new textbooks and studying the effect on individual student performance
  - Offering a new financial service to all residents in a village and studying the impact on micro enterprise outcomes
- In a clustered randomization the power of the study is coming partly from the number of individuals in the study, and partly from the number of clusters in the study

Cluster Randomized Experiments

- The estimate of treatment effect is $\widehat{\beta}$ in the regression

$$Y_{ij} = \alpha + \beta T_j + \omega_j + \varepsilon_{ij}$$

- σ^2 is the variance of the outcome (ε_{ij})
- τ^2 is the variance of the outcome (ω_j)
- p is the proportion of treated units
- n is the number of observations in each cluster
- J is the number of clusters
- The variance of $\widehat{\beta}$ is $\sigma_{\widehat{\beta}} = \frac{n\tau^2+\sigma^2}{p(1-p)nJ}$

Cluster Randomized Experiments

- Often, expressed using the intra-cluster correlation (ICC) $\equiv \frac{\tau^2}{\tau^2+\sigma^2}$

- The variance of $\widehat{\beta}$ is $V(\widehat{\beta}) = \sigma^2 \frac{\rho + \frac{(1-\rho)}{n}}{p(1-p)J}$

- The ICC can be obtained using *loneway* in stata

Minimum detectable effect

- The **minimum detectable effect** is given by

$$MDE = (t_{\frac{\alpha}{2}} + t_{1-\kappa})\sigma\sqrt{\frac{\rho + \frac{(1-\rho)}{n}}{p(1-p)J}}$$

Power Calculations Rules of Thumb

- For an individual-level experiment, 200-300 observations will typically be sufficient to detect a reasonable effect size

- For a clustered experiment, a low ICC (0.1) would need 50-100 clusters and > 5 observations per cluster to detect a moderate effect. As the ICC gets larger, the number of **clusters** has to go up

- For **very** complicated research designs, you can always use simulations to get the power of the design

Power, Blocking, and Stratification

Introduction

Statistical power

Blocking/Stratification

Relationship between Research Design and Analysis

Power, Blocking, and Stratification

Introduction

Statistical power

Blocking/Stratification

Relationship between Research Design and Analysis

Blocking/Stratification

- A significance of a test is the chance that you have significant imbalance between the treatment and the control for a given variable

- What if you have a variable that you want to ensure is balanced across treatment and control?

- This can easily be achieved by using this variable in the randomization

Blocking/Stratification

- Take a binary or categorical variable that describes the groups you are concerned about balance over (gender, occupational categories, geographical regions)
- Perform the randomization to assure that exactly a share p units is treated within each group
- The experiment is balanced across groups by definition
- This also implies that we have a replica of the experiment within each subgroup
    - You are in the best position to examine treatment effects by subgroup
- Analysis of a blocked randomization should include fixed effects for the blocks

Blocking/Stratification

- To conduct a blocked randomization:
    - Start with all the observations within which you want to randomize
    - Create a variable that identifies the blocks
    - Create a random number (using randomvar = runiform() command in Stata)
    - In Stata you will get a different time you run your do file unless you have set the seed. To do this, include the command ‘set seed madeupnumber’
    - Sort the data first by the group identifier, and then by the random number
    - Take the first fraction *p* of every group and assign to treatment

Blocking/Stratification

- Natural relationship between blocked or stratified designs and pre-commitment in experimentation
- When there is no effect, at the 5% level, 1/20 variables will be significantly different between treatment and control
- One solution is a pre-analysis plan that specifies the hypotheses you intend to test
- Another variant of this problem is looking for heterogeneity in treatment effects
- Signal that you are interested in examining a specific type of heterogeneity by blocking/stratifying on that characteristic

How to think through the way to randomize

- Kernan et al. (1999) summarize the potential advantages of stratifying:
  - Balance on variables correlated with the outcome of interest
  - Protecting against type I error (by reducing the chance of imbalance)
  - Facilitating sub-group analysis by assuring balance of treatment status for this subgroup
  - Protecting against “stratas” dropping-out of the experiment (still have a valid experiment for the other strata)
  - Increasing power, and therefore efficiency, by reducing the residual variance (but not always)

How to think through the way to randomize

- Trade off of blocking on more attributes in the randomization:

$$\frac{V(\beta_{\text{without controls}})}{V(\beta_{\text{with controls}})} = \frac{n - 2\sum \widehat{u}_i^2}{n - k - 2\widehat{\varepsilon}_i^2}$$

- $\widehat{\varepsilon}$ is the residual once the blocks fixed effects are included

- k is the number of degrees of freedom lost (number of blocks)

- $\widehat{u}$ is the residual when blocks are not included

How to think through the way to randomize

- Trade off of blocking on more attributes in the randomization:

$$\frac{V(\beta_{\text{without controls}})}{V(\beta_{\text{with controls}})} = \frac{n - 2\sum \widehat{u}_i^2}{n - k - 2\widehat{\varepsilon}_i^2}$$

- $\widehat{\varepsilon}$ is the residual once the blocks fixed effects are included
- k is the number of degrees of freedom lost (number of blocks)
- $\widehat{u}$ is the residual when blocks are not included
- Blocking on a completely irrelevant variable may decrease power

Re-randomization or ‘Big Stick’

- Write a loop to iterate the randomization many times, and then pick the ‘best’ randomization
- Two ways of doing this
  - Test for balance on a set of covariates and iterate until all p-values look good
  - Conduct the randomization X times and then pick the one that has the best balance
- There is no way to adjust the analysis of the experiment for the way the randomization was done
- Forced to do randomization inference
- Beware human error, and use simpler methods!

Power, Blocking, and Stratification

Introduction

Statistical power

Blocking/Stratification

Relationship between Research Design and Analysis

Power, Blocking, and Stratification

Introduction

Statistical power

Blocking/Stratification

Relationship between Research Design and Analysis

“As Ye Randomize, So Shall Ye Analyze”

- Bruhn and McKenzie provide good rules of thumb for how you’ll have to handle your data based on the way the randomization is done:
  - Cluster the standard errors in a regression on a Cluster Randomized Trial
  - Include fixed effects for the blocks used in randomization
  - No easy way to adjust regressions for re-randomization routines, which should make us leery of these. **Need to conduct randomization inference**
  - If you have a small sample and use re-randomization over a large number of draws the sample becomes almost deterministic
- Typical regression in Stata:
  reghdfe outcome treatment, absorb(strata) vce(cluster groups)

Do balance tables make any sense?

- Why would you test for the number of imbalances that occur when you know that the imbalances occur by random chance?

Do balance tables make any sense?

- Why would you test for the number of imbalances that occur when you know that the imbalances occur by random chance? Answer: you might have screwed something up!

- Testing balance on variables for which you forced balance through blocking/stratification/re- randomization is completely degenerate

- Hard to confirm that variables presented haven't been systematically chosen

- Should you adjust for imbalanced covariates?

Do balance tables make any sense?

- Why would you test for the number of imbalances that occur when you know that the imbalances occur by random chance? Answer: you might have screwed something up!
- Testing balance on variables for which you forced balance through blocking/stratification/re- randomization is completely degenerate
- Hard to confirm that variables presented haven't been systematically chosen
- Should you adjust for imbalanced covariates? Freedman says no

Do balance tables make any sense?

- Why would you test for the number of imbalances that occur when you know that the imbalances occur by random chance? Answer: you might have screwed something up!
- Testing balance on variables for which you forced balance through blocking/stratification/re- randomization is completely degenerate
- Hard to confirm that variables presented haven't been systematically chosen
- Should you adjust for imbalanced covariates? Freedman says no but it is difficult to avoid doing this once you've shown large imbalances on a critical covariate.
- **For better or worse, balance tests persist**