Lecture 14: Design and Analysis of Experiments

Jaakko Meriläinen

5304 Econometrics @ Stockholm School of Economics

Introduction

- In Lecture 1, we started with examples of how random allocation deals with the selection problem
- From Lectures 2 – 13, we kept the focus on causal effects but implicitly imposed that we did not have experimental variation
- The one theme running through has been selection bias and how to deal with that
- This lecture returns to the starting point – suppose we could design our own experiment!
  - How would you go about designing it?
  - What are design-related issues that you need to keep in mind in analysis?

Plan for Today

1. Introduction
2. Preliminaries before you go experimenting
3. Statistical power
4. Individual versus clustered trials
5. Baseline covariates and block randomization
6. Special considerations
   - Imperfect compliance
   - Two-stage designs
7. Example: School vouchers in India
8. Concluding remarks

You Are in Good Company!

The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2019

© Nobel Media. Photo: A. Mahmoud

Abhijit Banerjee

Prize share: 1/3

© Nobel Media. Photo: A. Mahmoud

Esther Duflo

Prize share: 1/3

© Nobel Media. Photo: A. Mahmoud

Michael Kremer

Prize share: 1/3

Defining the Research Question

- Designing a randomized control trial (RCT) offers the huge advantage of being able to control the variation in your independent variable
- Done right, you eliminate selection bias in expectation, but there are also challenges
- You need to, at the very beginning, make sure that:
  - your research question is precisely defined
  - your experiment does, in fact, map on to the question you want to answer
  - your design is likely to have sufficient statistical power
  - your design takes ethical considerations into account
  - your design is actually feasible
- These challenges are not specific to RCTs, what is specific is that you need to deal with them upfront
- If you mess up the design, your RCT might be unsalvageable!

Statistical Power

- Let us assume you do have your research question well thought-out and you want to test it with an RCT
- Eventually, we want to be able to test some hypothesis H_0 vs an alternative H_1
- There are two types of errors we can make:
  - **Type I error:** reject H_0 when H_0 is true
  - **Type II error:** fail to reject H_0 when H_0 if false
- In designing a trial, you need to keep both types of errors in mind!

Statistical Power

- The **significance level** we use for hypothesis tests gives the probability of a Type I error—typically, 5% i.e. 0.05 level of significance

- The **power of a test** is the probability of correctly rejecting the null hypothesis when the null hypothesis is false
  - Power = 1 - the probability that you make a Type II error
  - Common choices for power in the literature are 90% and 80%

- Having fixed these two, our usual questions are one of two types:
  - **How large a sample size** do I need to estimate effects of the size I expect to see?
  - Given the sample size available to me, what is the **minimum detectable effect** in the experiment?

Type I and II Errors (Marc Gurgand 2011)

		YOU CONCLUDE	
		Effective	*No Effect*
THE TRUTH	*Effective*		Type II Error 
	No Effect	Type I Error  (probability = significance level)	

Type I and II Errors (Marc Gurgand 2011)

		YOU CONCLUDE	
		Effective	*No Effect*
THE TRUTH	*Effective*	 (probability = power)	Type II Error 
	No Effect	Type I Error 	

Statistical Power

- Let us take the case of an individually-randomized binary treatment variable (T)
- You want eventually to run the following regression:

$$Y_i = a + \beta T_i + \epsilon_i$$

- Because of randomization, we know $E(\hat{\beta}) = \beta$ (i.e., $\hat{\beta}$ is an unbiased estimator)
- In large samples, the variance of $\hat{\beta}$ is $V(\hat{\beta}) = \frac{\sigma^2}{p(1-p)N}$
  - σ^2 is the variance of the outcome (Y_i); often, for power calculations, this is normalized to $\sigma^2 = 1$
  - p is the proportion of treated units
  - N is the number of observations

Minimum Detectable Effect

- With given power size (κ), significance level (α), sample size (N), and proportion of individuals treated (p), the **Minimum Detectable Effect (MDE)** is given by:

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

- Here $t_{\frac{\alpha}{2}}$ and $t_{1-\kappa}$ refer to relevant critical values from the t-distribution
  - With $\alpha = 0.05$, $t_{\frac{\alpha}{2}} = 1.96$
  - With $\alpha = 0.05$ and $\kappa = 0.8$, $t_{1-\kappa} = 0.84$
  - Implicitly assuming the same σ across treated and control groups

- So, with 5% level of significance and 80% power, the MDE is about 2.8 $\times$ the standard error of $\hat{\beta}$

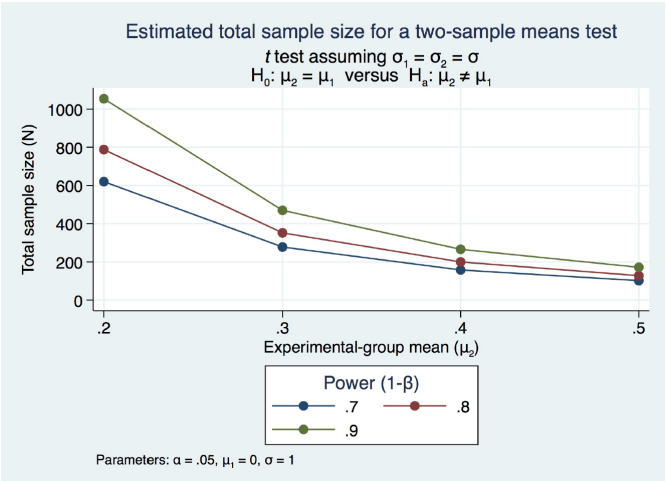
Minimum Detectable Effect

- Let us look at this formula for a bit

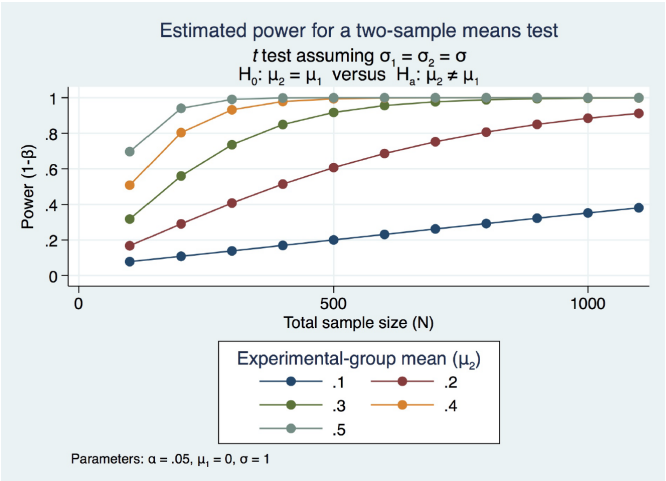
$$MDE = (t_{\frac{\alpha}{2}} + t_{1-\kappa})\sqrt{\frac{\sigma^2}{p(1-p)N}}$$

- We can plug in values to find answers we want—for instance:
  - Given $\alpha, \kappa, \sigma^2, p$, how large a sample do I need?
  - Given other parameters, what is the minimum size of effect I will be able to detect?
  - Given other parameters, what is the actual power (κ) of the experiment?

How Many People Do I Need?



How Much (Statistical) Power Do I Actually Have?



Deciding How Large a Trial Should Be?

- Note: these numbers are needed before you start a trial
  - You do not actually know the right means, SD, or the likely effect size
  - These tend to come from previous studies or pilots
- **Do not run underpowered trials**
  - Effects may go undetected, even if sizable
  - The estimates will be very noisy, we may not learn very much at all
- **What is the smallest meaningful size?**
  - The MDE should be closely linked to what you think is economically meaningful
- **What is feasible for you to actually set up and evaluate?**
  - Not all aspects will be under your direct control

What Level Should I Randomize at?

- So far, we have only been talking about individually-randomized RCTs
- But sometimes, the unit of randomization needs to be higher than the individual
- Two prominent cases:
  - The treatment is naturally clustered
    - E.g. you upgrade a school, all pupils are affected
  - The treatment is likely to have spillovers across individuals
    - E.g. if I gave cash transfers to some of you and thought you might share

Power Analysis in a Cluster Randomized Trial

- With cluster randomized trials, the residual variation is no longer independent across observations
- The model could be written as:

$$Y_{ic} = a + \beta T + u_c + \epsilon_{ic}$$

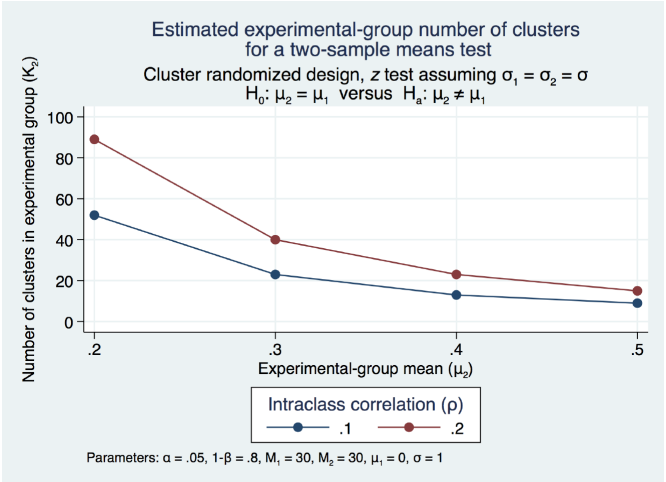
where u_c is the group component of the residual variation

- Then MDE with 5% level of significance and 80% power is

$$MDE = (1.96 + 0.84)\sqrt{\frac{1}{p(1-p)}(\frac{V(u_c)}{N_c} + \frac{V(\epsilon_{ic})}{N})}$$

- 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
- The higher the ratio of the variance within-cluster to the total variance, the larger the MDE

Power Analysis in a Cluster Randomized Trial



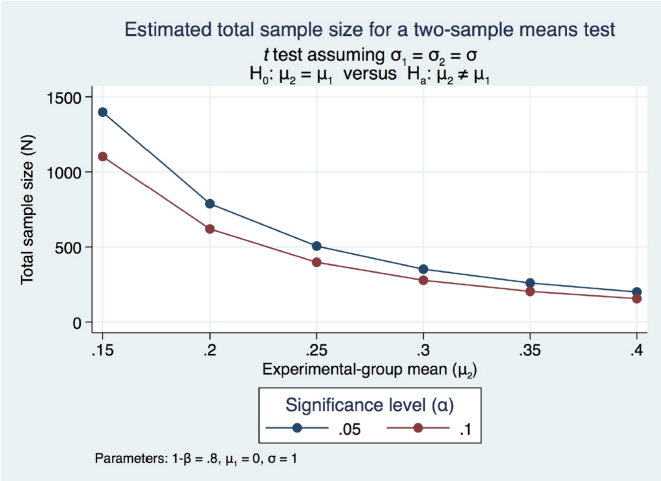
Power Analysis in a Cluster Randomized Trial

- The loss in precision in moving from an individually-randomized to a cluster-randomized trial can be **very large**

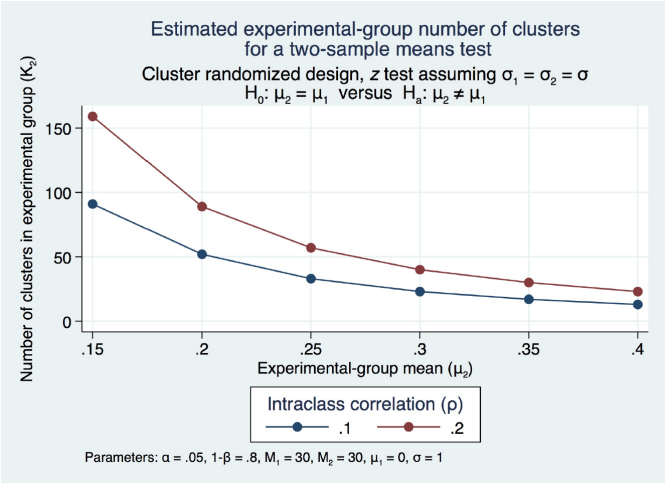
- Let us take a concrete example:
  - Suppose you wanted to roll out an individual instruction program to see the effect on student grades
  - You care about effects larger than 0.15 SD
  - You have two potential models:
    - Individually randomize students to treatment and control groups
    - Randomize classrooms into treatment and control groups
    - Assume you will test 30 students in each class

  - What are the implications for how many students you need to cover?

Required Sample in an Individually-Randomized Trial



Required Sample in a Cluster-Randomized Trial



Other Ways of Improving Statistical Precision

- So far, we have paid attention primarily to the role of N in improving statistical power
- But that is not the only way
- We can also reduce the variance (σ) in the dependent variable...
  - by controlling for baseline covariates
  - by better measurement
  - by stratifying randomization

Blocking/Stratification

- Suppose this is a binary or categorical variable
  - What you need to ensure is that the probability of treatment (p) is exactly the same across groups
  - The simplest example is where you divide the clusters into pairs and randomize one into treatment
- 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

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

RCTs with Imperfect Compliance

- So far, we have been in the setting where we have assumed that all clusters/individuals assigned to treatment are in fact treated, none of the controls are
  - But we have discussed earlier that this is not always the case
  - Some units assigned to treatment may go untreated
  - Some units assigned to control may get access to treatment

- An example is an "encouragement design" where treatment group are given an option/inducement to participate, but may choose not to

- What can we do then?
  - As we discussed in the IV section, the assigned treatment can be used as an IV for the actual treatment

- What does this mean for power calculations?
  - You should run these with the expected ITT effect, not the (L)ATE

SUTVA Violations

- **"Spillovers"**
  - Imagine for example a voucher scheme for students (peer effects?)
  - Or the effect of a scholarship for one child in the household (there is a HH budget constraint)
  - Or providing information to some students in a class

- **"General Equilibrium Effects"**
  - The employment, price etc. effects of a large-scale (universal) basic income experiment versus the effects if this was done in a smaller scale

Two-Stage Designs

- One possible way to deal with violations like the above is to randomize at two stages:
  - First, across the level at which spillovers/GE effects are expected to occur
  - Second, across individuals within these clusters
- The individuals in the untreated cluster provide a "pure" control group
- Comparing average outcomes across treated and control clusters provides a clean experiment of the cluster-level treatment effects
- The difference in outcomes between the (randomly assigned) treated and untreated individuals in the treated clusters shows the GE effects

Spillovers in Cash Transfers (Angelucci and De Giorgi 2009)

Indirect Effects of an Aid Program: How Do Cash Transfers Affect Ineligibles' Consumption?

By Manuela Angelucci and Giacomo De Giorgi*

Cash transfers to eligible households indirectly increase the consumption of ineligible households living in the same villages. This effect operates through insurance and credit markets: ineligible households benefit from the transfers by receiving more gifts and loans and by reducing their savings. Thus, the transfers benefit the local economy at large; looking only at the effect on the treated underestimates the impact. One should analyze the effects of this class of programs on the entire local economy, rather than on the treated only, and use a village-level randomization, rather than selecting treatment and control subjects from the same community.

General Equilibrium Effects: Price Effects of Cash vs. In-Kind Transfers (Cunha et al. 2018)

Abstract

This article examines the effect of cash versus in-kind transfers on local prices. Both types of transfers increase the demand for normal goods; in-kind transfers also increase supply in recipient communities, which could lead to lower prices than under cash transfers. We test and confirm this prediction using a programme in Mexico that randomly assigned villages to receive boxes of food (trucked into the village), equivalently-valued cash transfers, or no transfers. We find that prices are significantly lower under in-kind transfers compared to cash transfers; relative to the control group, in-kind transfers cause a 4% fall in prices while cash transfers cause a positive but negligible increase in prices. In the more economically developed villages in the sample, households' purchasing power is only modestly affected by these price effects. In the less developed villages, the price effects are much larger in magnitude, which we show is due to these villages being less tied to the outside economy and having less competition among local suppliers.

Comparing Government and Private Schools (Muralidharan and Sundaraman 2015)

School and Teacher Characteristics

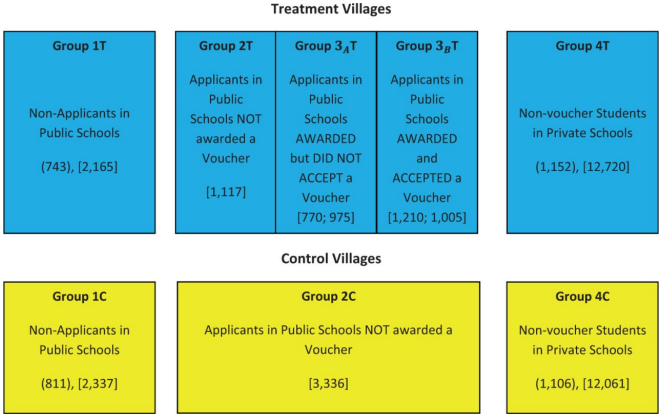
	(1) Private schools	(2) Public schools	(3) Difference
Panel A: School characteristics			
Total enrollment	296.21	74.04	222.17***
Total working days	229.81	218.66	11.15***
Pupil-teacher ratio	17.62	25.28	−7.67***
Drinking water available	0.99	0.92	0.07***
Functional toilets	0.86	0.68	0.18***
Separate functional toilets for girls	0.77	0.40	0.37***
Functional electricity	0.88	0.61	0.28***
Functional computers	0.52	0.05	0.48***
Functional library	0.80	0.97	−0.18***
Functional radio	0.13	0.81	−0.68***
Observations	289	346	
Panel B: Teacher characteristics			
Male	0.24	0.46	−0.21***
Age	27.58	40.00	−12.42***
Years of teaching	5.14	14.96	−9.82***
Completed at least college or masters	0.69	0.88	−0.19***
Teacher training completed	0.34	0.99	−0.65***
Come from the same village	0.44	0.13	0.32***
Current gross salary per month (Rs)	2,606.66	14,285.94	−11,679.27***
Observations	2,000	1,358	
Panel C: School expenditures			
Annual cost per child (Rs/child)	1,848.88	8,390.00	−6,542***
Observations	211	325	

Comparing Government and Private Schools (Muralidharan and Sundaraman 2015)

TEACHER AND SCHOOL EFFORT

	(1) Private schools	(2) Public schools	(3) Difference
Panel A: Measures of classroom activity			
Class is engaged in active teaching	0.51	0.34	0.17***
A teacher is present in class	0.97	0.92	0.048***
Teacher is effective in teaching and maintaining discipline	0.50	0.36	0.14***
Teacher has complete control over class	0.69	0.41	0.28***
Teachers teaching mutliple classes at the same time	0.24	0.79	−0.55***
Observations	2,738	2,784	
Panel B: Measures of teacher activity			
Teacher is absent	0.09	0.24	−0.15***
Teacher is actively teaching	0.50	0.35	0.15***
Teacher is in school and not teaching	0.01	0.03	−0.02***
Observations	6,577	5,552	
Panel C: Measures of school hygiene			
Flies heavily present on premises of the school	0.14	0.19	−0.05**
Stagnant water present on premises of the school	0.18	0.28	−0.10***
Garbage dumped on premises of the school	0.33	0.44	−0.11***
Observations	426	614	

Experimental Design



Results on Test Scores

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
	Year 2 assessments				Year 4 assessments						
	Telugu score	Math score	English score	Combined across tests	Telugu score	Math score	English score	EVS score	Combined across tests excluding Hindi	Hindi score	Combined a
Panel A: Impact of winning a voucher (intention to treat effects)											
Offered voucher	−0.079	−0.053	0.185**	0.016	−0.017	−0.031	0.116*	0.083	0.036	0.545***	0.133***
	(0.055)	(0.065)	(0.079)	(0.061)	(0.051)	(0.052)	(0.070)	(0.060)	(0.048)	(0.068)	(0.045)
Total observations	4,620	4,620	4,525	13,765	4,385	4,385	4,217	4,243	17,230	1,696	18,926
Treatment observations	1,778	1,778	1,738	5,294	1,674	1,675	1,607	1,628	6,584	867	7,451
Control observations	2,842	2,842	2,787	8,471	2,711	2,710	2,610	2,615	10,646	829	11,475

Putting Results in Context: School Time-Tables

TABLE VII

SCHOOL TIME USE: INSTRUCTIONAL TIME BY SUBJECT (MINUTES PER WEEK)

	(1) Private schools	(2) Public schools	(3) Difference
Telugu	307.72	511.52	−203.81***
	(6.36)	(3.60)	(6.99)
Math	339.75	500.69	−160.94***
	(7.50)	(3.36)	(8.63)
English	322.68	235.52	87.17***
	(7.96)	(5.39)	(9.69)
Social studies	239.21	173.24	65.96***
	(6.29)	(6.89)	(9.84)
Science	205.52	104.58	100.94***
	(9.09)	(5.78)	(9.44)
Hindi	215.78	0.01	215.77***
	(6.08)	(0.89)	(6.41)
Moral science	16.85	20.11	−3.26
	(4.82)	(3.20)	(5.56)
Computer use	46.7	0.51	46.19***
	(6.50)	(1.02)	(6.80)
Other	311.66	250.29	61.37***
	(14.55)	(6.70)	(16.20)
Total instructional time	2,005.87	1,796.47	209.4***
	(13.73)	(6.86)	(14.46)
Break	461	473.18	−12.18
	(9.14)	(3.05)	(10.58)
Total school time	2,466.87	2,269.65	197.22***
	(17.46)	(8.25)	(19.79)
Observations	325	200	

Some Final Thoughts

- The design of experiments is partly an art
  - There are principles but also a lot of educated guesses and some craftsmanship
- The actual design is almost always some combination of
  - power considerations
  - logistical considerations
  - budgets
- In addition to these, there are important considerations of ethics
  - All trials run by e.g. JPAL and IPA require approval from an ethics board
  - Governed by well-established principles for Human Subjects Research
  - This is not true of industry trials
  - In your future lives, e.g. in Facebook or McKinsey, do no harm knowingly!

References

- Angelucci, M., & De Giorgi, G. (2009). Indirect effects of an aid program: how do cash transfers affect ineligibles' consumption?. American Economic Review, 99(1), 486-508.
- Cunha, J. M., De Giorgi, G., & Jayachandran, S. (2018). The price effects of cash versus in-kind transfers. The Review of Economic Studies, 86(1), 240-281.
- Duflo, Esther, Rachel Glennerster, and Michael Kremer: "Using Randomization in Development Economics Research: A Toolkit" in Handbook of Development Economics, Vol. 4, 2007
- Muralidharan, K., & Sundararaman, V. (2015). The aggregate effect of school choice: Evidence from a two-stage experiment in India. The Quarterly Journal of Economics, 130(3), 1011-1066.