BREAD-IGC-ISI Summer School in Development Economics

Prof. Karthik Muralidharan

Lecture 3 (Experimental Methods)

Lecture Plan

1. Introduction to Randomized Experiments (20)
2. Designing and Implementing an Experimental Evaluation (30)
3. Strategies for Introducing Randomization (15)
4. Outcome measures (20)
5. Power calculations and sample size (15)
6. Managing threats (next lecture)
7. Data analysis (next lecture)

Lecture Outline

- **Introduction to Randomized Experiments**

Improving Education – Flip Charts

- An officer in the Kenyan education ministry comes to the minister and suggests that children learn better with visual aids and pictures than with text books – also much more cost effective because only need one per classroom

- He has found a producer who can make large flip charts with pre-printed pictures that are mounted on an easel and can be used in a 'show and tell' fashion

- He strongly recommends that the Ministry of Education procure flip charts and distribute them to all government-run schools as a cheap way of improving quality of education

- You are the advisor to the minister – what do you tell her to do?

Impact Evaluation - Flip Charts (1 of 2)

- The relationships that you want to estimate is:
- A = h(**School Quality**, **Child**, **HH**, **Peers**)
- "School Quality" is a composite measure that can be affected in several ways and providing additional inputs to schools is one way of doing so
- The flip-charts are an example of this:
- The relationship that the officer has estimated is:

$$Norm\_Test\_Score_{ijk} = \beta_0 + \beta_1(Flip\_Chart_k) + \mu_{ijk}$$

- He has estimated that the value of $\boldsymbol{\beta_1}$ is 0.2
- What do you make of this?

Impact Evaluation Flip Charts (2 of 2)

- The previous result is subject to several omitted variable problems
- These are all forms of 'selection' bias – go back to the question of:
  - "Why is there variation in the right hand side of the regression?"
  - In other words, why is there variation in who gets the treatment?
- However, consider what was actually done here:
- A randomized experiment with a subset of pre-defined population getting the flip charts by lottery:

$$Norm\_Test\_Score_{ijk} = \beta_0 + \beta_1(Flip\_Chart_k) + \mu_{ijk}$$

- The estimating equation is the same as before, so what is the difference here?
- Now, we find that there was no effect of the flip charts at all

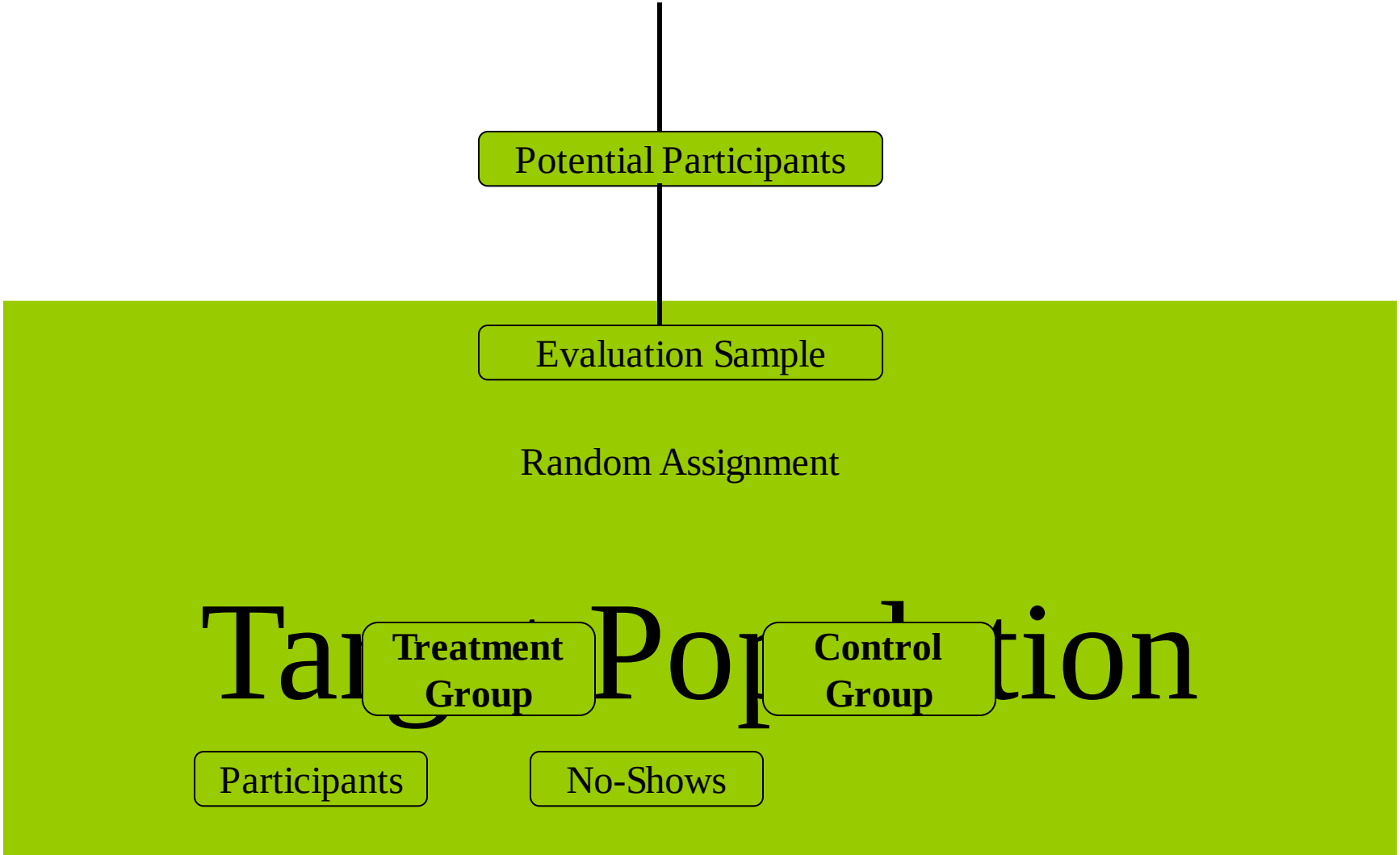
Selection and Omitted Variables Bias

- The correlation between infrastructure and absence could be driven by an omitted variable
  - Examples include community, political leaders who care about education
  - This could both improve infrastructure and reduce teacher absence
  - So the correlation is not the same as causation
- Connection between omitted variable bias and selection bias
  - Consider the “treatment” to be provision of infrastructure
  - Policy research question is what is the impact of this “treatment” on teacher absence
  - The omitted variable can be thought of as causing the selection bias (and also being correlated with the outcome!)

Randomization Solves the OVB Problem

- Suppose x% of a group of eligible individuals are randomly `treated' to a program (without regard to their characteristics).
  - If successfully randomized, individuals assigned to the treatment and control groups differ only in their exposure to the treatment.
  - Implies that the distribution of both observable and unobservable characteristics in the treatment and control groups are statistically identical.
  - Had no individual been exposed to the program, the average outcomes of treatment and control group would have been the same, and there would be no omitted variable or selection bias!

Basic Setup of a Randomized Evaluation



Key advantage of experiments

Because members of the groups (treatment and control) do not differ systematically at the outset of the experiment, any difference that subsequently arises between them can be attributed to the treatment rather than to other factors

Validity

- We care about two types of validity
  - **Internal validity**: relates to ability to draw causal inference, i.e. can we attribute our impact estimates **in the context we are studying** to the program and not to something else
  - **External validity**: relates to ability to generalize to other settings of interest, i.e. can we generalize our impact estimates from this program/context to other populations, time periods, countries, etc?
- Random assignment versus random sampling
  - Random assignment helps with internal validity
  - Random sampling helps with external validity

Other advantages of experiments

- Relative to results from non-experimental studies, results from experiments are:
  - Less subject to methodological debates
  - Easier to convey
  - More likely to be convincing to program funders and/or policymakers
- Striving for an experimental evaluation of all new programs is an excellent disciplining device for ensuring that you think about program impact (or lack thereof) in an objective way

Limitations of experiments

- Despite great methodological advantage of experiments, they are also potentially subject to threats to their validity
  - **Internal Validity**
    - Hawthorne Effects, survey non-response, no-shows, crossovers, implementation issues, etc.
    - Substitution Bias
  - **External Validity**
    - Are the results generalizable to the population of interest?
    - Randomization Bias
- But these threats usually also affect the validity of many non-experimental studies
- Randomization is not a panacea – but it is often the 'cleanest' way to achieve causal inference

Other limitations of experiments

- Not all questions are amenable to study by experiments
  - Complex, compound interventions
  - Reduced form estimates vs. fundamental parameters (out of sample predictions)
- Costs
  - Larger than that of observational studies
  - But the gains in 'discipline' are often worth it
- Partial equilibrium
  - Contract teachers, performance-pay, vouchers
- Ethical Issues
  - How can we deny a program to anyone?
- Can often provide 'black box' treatment effects without an understanding of process/mechanisms

Summary

- While Randomized control trials have important limitations, they still offer the most promise for rigorous impact evaluation of programs
  - Considered the “gold standard” for a good reason
- While not all questions are amenable to experimental evaluation, the scope for randomized evaluations is much higher than we think
  - Education, health, finance, governance, etc.
- The case for experimental evaluations of programs and policies is particularly high in developing countries
  - Cost of spending on ‘ineffective’ policies is much higher
  - Resource constraints mean that a ‘phase in’ design is often adopted anyway – a lottery is a natural way to do this

Lecture Outline

- **Designing and Implementing an Experimental Evaluation**

Key Issues in an Experimental Study

1. Defining the Research Question
2. Defining the features of the program that are amenable to experimental or quasi-experimental investigation
3. Deciding on the randomization design
4. Building the Research Team; division of labor & relationships with the field staff.
5. A general timeline of the baseline, intervention, followup, and results dissemination.

What is the Research Question?

- By far the most important thing to have clarity on. Types of experimental studies include:
- Experiments to answer a directly relevant policy question: "What is the impact of a particular program"
- Using experiments to understand mechanisms of program impact: "How did behavior change as a result of X"
- Using experiments to test theory:
  - Does theoretically predicted heterogeneity show up in the data
  - Testing for theoretically-predicted negative/side effects
- Key is to think really hard upfront and try to design the experiment accordingly (several examples follow)

Unit of randomization

- Randomizing at the individual level
- Randomizing at the group level
  - School
  - Community
  - Health center
  - Sub-district
  - District
- How do you tell which level to randomize at?

Unit of randomization

- Individual randomization gives you a bigger sample size at lower cost
- May be politically difficult to have unequal treatment within a community
- Program can only be implemented at a certain level
- Spillovers
- Power
- Depends mainly on research question
- Examples

AP RESt Design Overview

	INCENTIVES (Conditional on Improvement in Student Learning)			
INPUTS (Unconditional)		NONE	GROUP BONUS	INDIVIDUAL BONUS
	NONE	CONTROL (100 Schools)	100 Schools	100 Schools
	EXTRA PARA TEACHER	100 Schools		
	EXTRA BLOCK GRANT	100 Schools		

Kenya Extra Teacher Project (ETP)

- Many countries have large class sizes and have recruited local or informal teachers to help teach. There are many important questions this raises:
  - What is the impact of smaller class sizes
  - How do local/informal teachers compare to more experienced but less accountable govt. teachers?
  - How important for learning is the peer group?
  - Is streaming/tracking beneficial?
- How can you design an evaluation that examines all these questions?
- Lets take them one by one.

Mapping Questions into Design: ETP

- Program only introduced for grade 1
- Some schools receive extra teachers some do not (answers question 1)
- To test the effect of teacher, randomize students between govt. vs. informal teacher (question 2)
- Some schools are tracked some are not (selected randomly)
- Which stream gets informal teacher vs. govt teacher is randomized (i.e. randomize teacher allocation) question 3

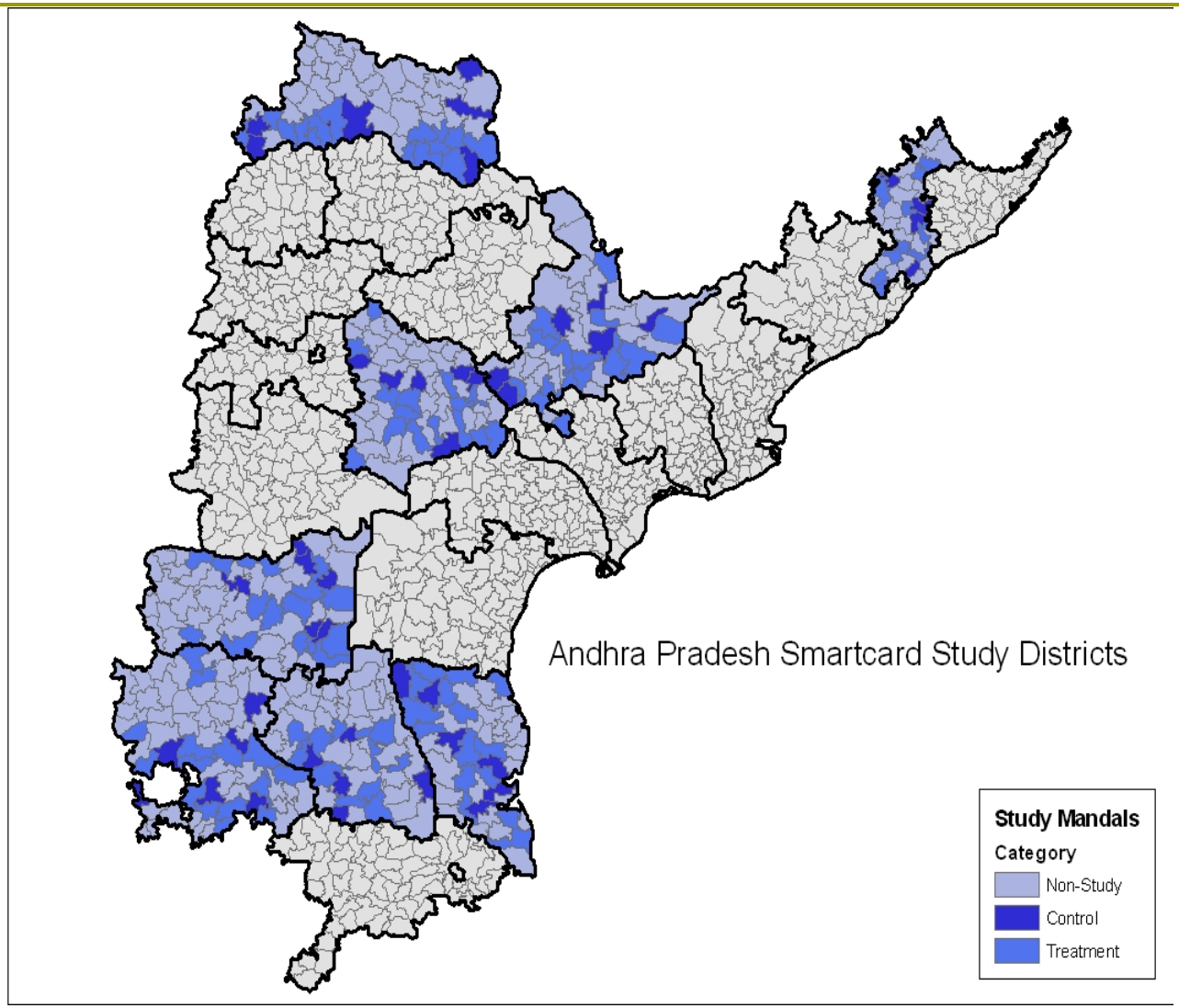
ETP – Design Matrix

Schools	Classes	Reduced Class Size	Pupil's assignment		Teacher	
			Random	Ability	Regular	Extra
Comparison (110 schools)	Group 1				X	
Treatment (220 schools)	Group 2A	X	X		X	
	Group 2B					X
	Group 3A	X		X	X	
	Group 3B					X

Andhra Pradesh Smartcard Impact Evaluation Study

- **Research Question:** What is the impact of using biometric smartcard payments in NREGA on leakage, beneficiary welfare, and volume of work
  - Have to randomize at the block/mandal level (lowest unit of program implementation)
  - Mandals in 8 districts randomized into treatment (Phase 1), non-study (Phase 2), control categories (Phase 3)
  - Baseline and endline surveys conducted to measure leakage in the system; transaction costs incurred in accessing payments; and welfare outcomes for beneficiaries (positive and negative)
  - MoU with GoAP calls for complete enrollment and fully carded payments in Phase 1 and 2 prior to any carded payments in Phase 3
  - 45 control & 112 treatment mandals across 8 districts

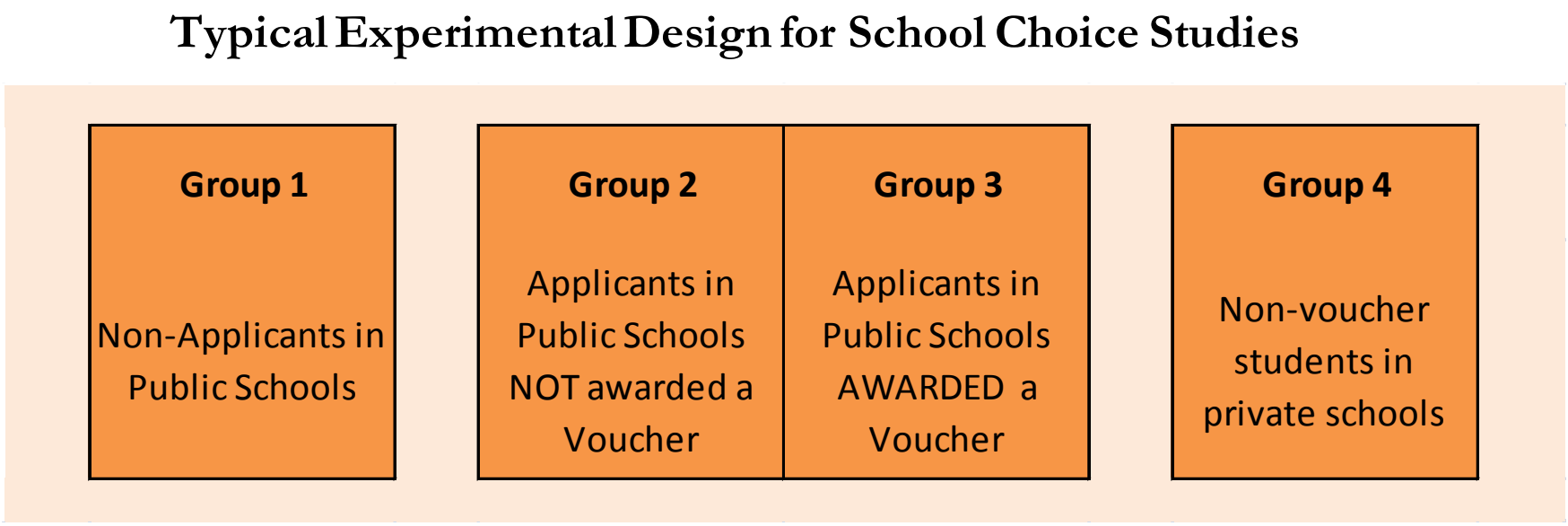
AP Smartcard Project Roll-out Map



AP Smartcard Project: Successful Randomization

Variable	Treatment Mean	Control Mean	Difference (p-value)
Population	43,846	43,807	38 (0.99)
Pensions Per Capita	0.12	0.12	0.00 (0.68)
Jobcards Per Capita	0.54	0.55	-0.01 (0.86)
Literacy Rate	0.45	0.45	0.00 (0.95)
SC Proportion	0.19	0.19	-0.00 (0.88)
ST Proportion	0.10	0.10	0.00 (0.98)
Proportion of Population Working	0.53	0.52	0.01 (0.47)
Proportion Male	0.51	0.51	0.00 (0.94)
Proportion of Pensioners under Indiramma	0.70	0.69	0.01 (0.52)
Proportion of Old Age Pensions	0.60	0.62	-0.02 (0.20)
Proportion of Weaver Pensions	0.01	0.01	-0.00 (0.64)
Proportion of Disabled Pensions	0.13	0.13	0.00 (0.76)
Proportion of Widow Pensions	0.26	0.25	0.02 (0.13)

Typical Design for School Choice Studies



AP School Choice Design

Design of the AP School Choice Project

Treatment Villages

Group 1T	Group 2T	Group 3T	Group 4T
Non-Applicants in Public Schools	Applicants in Public Schools NOT awarded a Voucher	Applicants in Public Schools AWARDED a Voucher	Non-voucher students in private schools

Control Villages

Group 1C	Group 2C	Group 3C	Group 4C
Non-Applicants in Public Schools	Applicants in Public Schools NOT awarded a Voucher	Does not exist	Non-voucher students in private schools

Anatomy of an Experimental Evaluation

- Planning
  1. Identify problem and proposed solution
  2. Identify key players
  3. Identify key operations questions to include in study
  4. Design randomization strategy
  5. Define data collection plan

- Pilot

- Implementation
  1. Identify “target” population (sample frame), and baseline data if applicable
  2. Randomize
  3. Implement intervention to treatment group
  4. Measure outcomes
  5. Repeat if possible!

Lecture Outline

- **Strategies for Introducing Randomization into evaluation designs**

Randomization Designs

- Lottery design
- Phase-in design
- Encouragement design
- Multiple treatments
- Rotation design (e.g., Balsakhi case)
  - Note: These are not mutually exclusive.

Lottery design

- Randomly choose from applicant pool
- Use when there is no good reason to choose among a subset of applicants and there is a budget constraint
- Lottery vs. targeting?
  - Lottery is often perceived as fair
  - High level of transparency possible
- Often politically more feasible
  - Risk of ex-post compensation for losers during the period of the study
- Can combine ‘screening’ for eligibility/need and a lottery (so lottery is conducted after screening)
  - Credit scoring (Karlan, Zinman)

Screening Plus Lottery

- Screening Rule
  - What are they screening for?
  - What elements are essential?
  - What elements are arbitrary?

- Example: Training program
  - 2000 candidates
  - 1000 fit criteria such as poor enough to be needy, qualified enough to take advantage of training
  - Only 500 slots exist
  - NGO willing to allocate slots by lottery within this group of 1000

- Lesson: Many parts of a selection rule are designed simply to weed out candidates
  - Example – teacher training grants in Indonesia

Phase-in Design

- In this design everyone is told that they will end up with the same outcome but some later than others.
- Use the fact that the program going to expand
- Example:
  - In 5 years, 500 schools will be covered.
- What determines which schools or districts will be covered in which year?
  - Some choices may be based on need, potential impact, etc.
  - Some choices largely arbitrary
- We can therefore choose who gets program early at random
  - Do this on population about which choices are arbitrary
  - AP Smartcards; Indonesia Teacher Training

Encouragement design

- In this design everyone is immediately eligible to receive the program—there is enough funding to cover all of them
- However not all of them will necessarily take it up
- Pick some people at random and encourage them to use program
- Use non-encouraged as control
- RSBY example
- What population is this a treatment effect on?
- Key idea is that the randomly-assigned "encouragement" is a valid "instrument" for the take up of the treatment of interest
  - Connect to earlier lecture on ATE, ITE, TET
  - As with any IV, the estimate here is a LATE

Encouragement

- What is an “encouragement”?
  - Something that makes some folks (treatment group) more likely to join than others (control group)
- Key criteria:
  - Should not itself be a “treatment”
    - Bad idea: intense entrepreneurship training program, as encouragement for credit
    - Good idea: marketing to some but not everyone; make sure the marketing isn’t overly informative
  - Think about who responds to the encouragement. Are they different?

Multiple Treatments

- Sometimes you are not sure what intervention to implement
- You can then randomize different interventions with different (randomly chosen) populations
- Does this teach us about the benefit of any one intervention?
- Advantage: win-win for operations, can help answer questions for them, beyond simple “impact”!

Convincing Partner Organizations

- But I already know the answer…
- (Between the lines): But I do not want to risk learning that we do not have impact.
- Finding the right & willing partner is key to making an impact evaluation work (we can talk more about this after class)

Lecture Outline

- **Measuring Outcomes**

What outcomes do we measure?

- The entire point of a program or policy is usually to have an impact on some outcome!
- Start by being clear about what need the program is looking to fill
- Be clear as to what outcomes your program is trying to impact and what the pathways to that impact are likely to be
- Collect data on both outcome and process variables to the extent possible

Example – Teacher Performance Pay

- Does teacher performance-pay improve test scores?
- What, if any, are the negative consequences?
- How do group and individual incentives compare?
- What is the impact of measurement and feedback?
- How does teacher behavior change?
- How cost effective is the incentive program?
- How will teachers respond to the idea?

Drawing the Chain of Causality

- We want to draw the link

  Intervention → intermediary variables → final outcome

- What are the intermediate variables through which the intervention is affecting the final outcome of interest?
- How are they likely to be affected?

Defining and measuring Intermediate outcomes will enrich our understanding of the program, reinforce our conclusions, and make it easier to draw general lessons

The Possible Effects

- Consider the teacher performance-pay program
- What process variables might it have an impact on?
  - Teacher attendance/motivation
  - Teacher effort
  - Student attendance/motivation
  - Teacher cooperation
  - Attention to non-incentive subjects
  - Cheating!
- What outcome variables might it have an impact on?
  - Student test scores in math/language
  - Scores on mechanical/conceptual; incentive/non-incentive
  - Student attrition
- In many cases, the best studies show some creativity or ingenuity in defining/measuring process/outcome variables

Estimating Production Function Versus Policy Parameters

- What is the difference?
- Example of a production function:
- A = h(**School Quality**, **Child**, **HH**, **Peers**)
- A production function parameter estimates how A would change when you change an input, holding everything else constant
- A policy parameter estimates the "total" effect on A when you change an input after all other inputs have been re-optimized
- Example with school grants in India
- Implications for data collection?

Data Collection

- Suppose budget was not an issue
- Should we go out, hire a large team of investigators, collect data on every single of these indicators in treatment and control schools, and see what happens? Pros/Cons?
- More data is usually good!
- But indiscriminate collection of data without a clear idea of what it is for can lead to:
  - Wasting money (usually this is the binding constraint!!)
  - Diluting effort of field enumerators and reducing quality
  - Survey fatigue if survey instruments are too long leading to perfunctory answers
  - Higher rates of rejection of surveys – leading to a more biased sample of respondents

In the Real World (Budget Constraints!)

- Collecting data is expensive, and we may have to make choices.
- If we had to choose one variable to determine whether the program made a difference, what would it be?
- You have computed the minimum budget required to conduct learning assessments of children in the treatment and control schools.
- What else can we find out with the same activities, without spending much extra money?
- If you had a slightly larger budget, what is the next set of variables you would focus on?

Data Collection

- Key point to keep in mind is that you don't have to collect data from EVERY subject in your study!
- A representative random sample is enough
- Often better to get more detailed answers from fewer respondents
- Key determinant of data collection strategy is fixed versus variable costs of additional data collection
- Examples:
  - Learning assessments versus household surveys
  - In some cases, the treatment may require that all subjects get measured (performance pay example)

Analyzing Multiple Outcomes

- Suppose you collect data on 20 measures of teacher behavior
  - You find that teachers assign significantly more homework in incentive schools
  - But students ask significantly fewer questions in incentive schools
  - There is no significant difference in the other 18 variables
- What do you make of these results?
- Hypotheses need to be designed in advance – else results can be subject to data mining

Stating Hypotheses Upfront

- What is the main hypothesis in your proposal?
- What is the key variable of interest (on which you will declare success or failure)?
- What are the intermediate outcomes that you are planning to collect data on?
  - What is your hypothesis on the channels by which your program may have an impact?
  - What are the negative outcomes you should check for?
  - Do you need to consider re-optimization by other agents?
- Heterogeneous treatment effects
  - This is usually of great interest
  - Same principles apply!

Data Collection and Management

- A common mistake is to monitor the 'treatment' units more carefully and collect better data there
  - This may be a program feature (cameras example)
- This can be a problem. Why?
  - Differences in data quality across treatments can produce sources of bias (cameras example)
  - Hawthorne effects
- Also, have to worry about differential attrition.
  - This can be a pretty big source of bias. Why?
- Solution is to track attrition through all stages of the project, and ideally to have a random subset of subjects who are always tracked down!

Lecture Outline

Power and Sample Size Calculations

Planning Sample Size for Evaluation

- General question:

  How large does the sample need to be to credibly detect a given effect size?

- Relevant factors:
  - Expected effect size
  - Variability in the outcomes of interest
  - Design of the experiment: stratification, control variables, baseline data, group v. individual level randomization

Basic set up

- At the end of an experiment, we will compare the outcome of interest in the treatment and the comparison groups.
- We are interested in the difference:

  Mean in treatment - Mean in control

  = Effect size
- This can also be obtained through a regression of Y on T:

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

  β = Effect size.

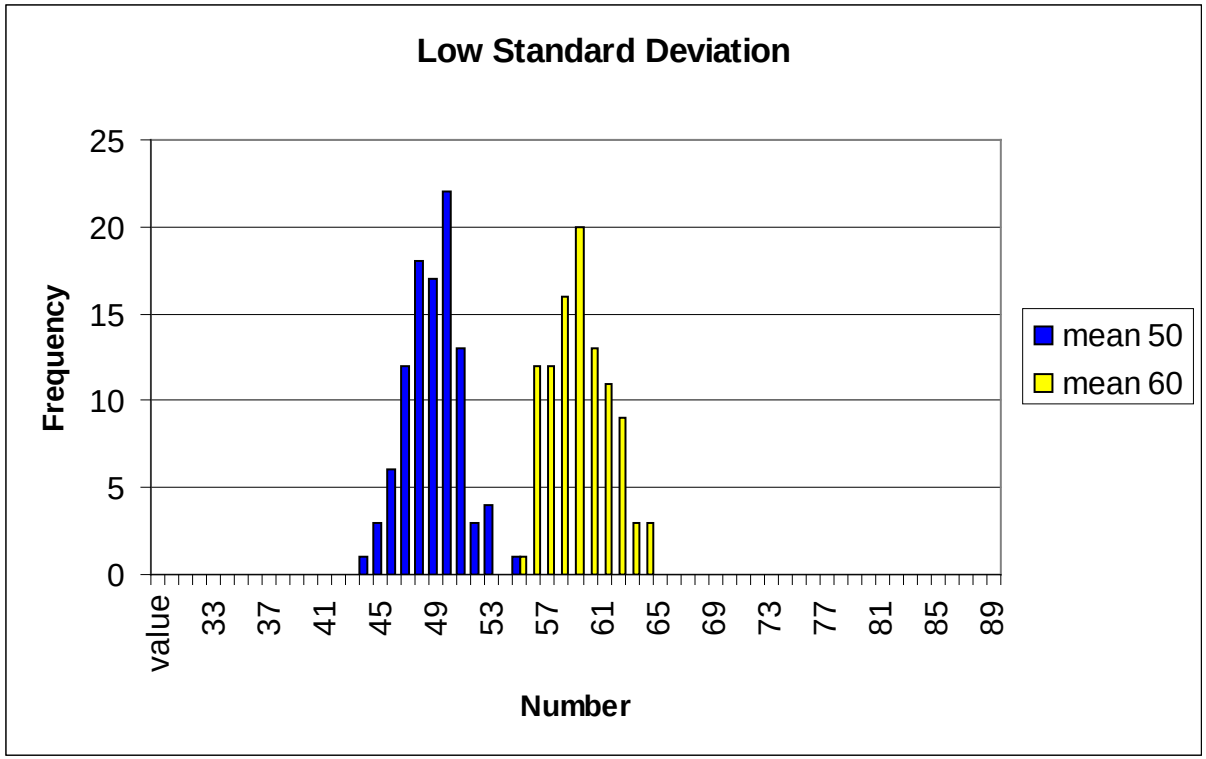
Estimation

We estimate the mean in our sample, for example by computing the sample average

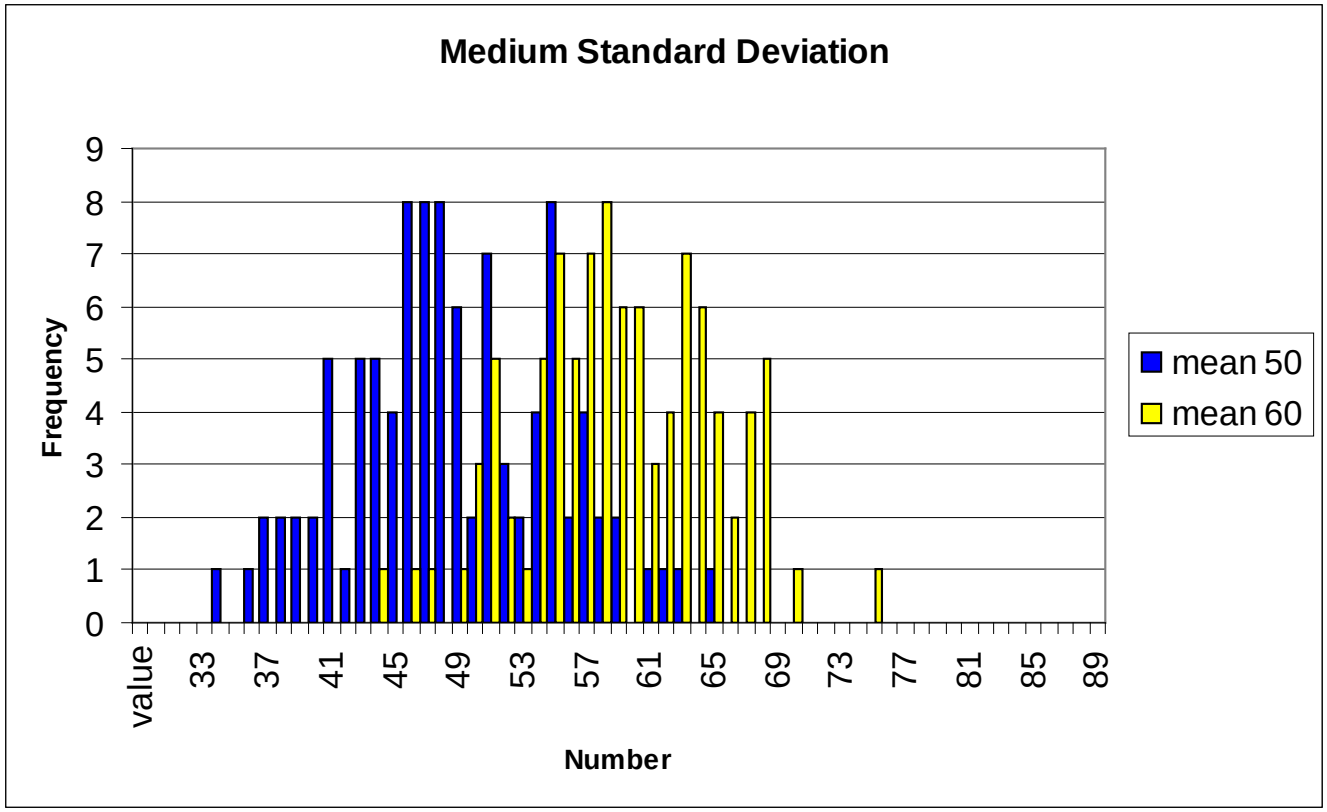
$$\widehat{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

If there is a lot of variation in the sample, the averages will be imprecise, so even if the real treatment effect is the same, it will be harder to detect.

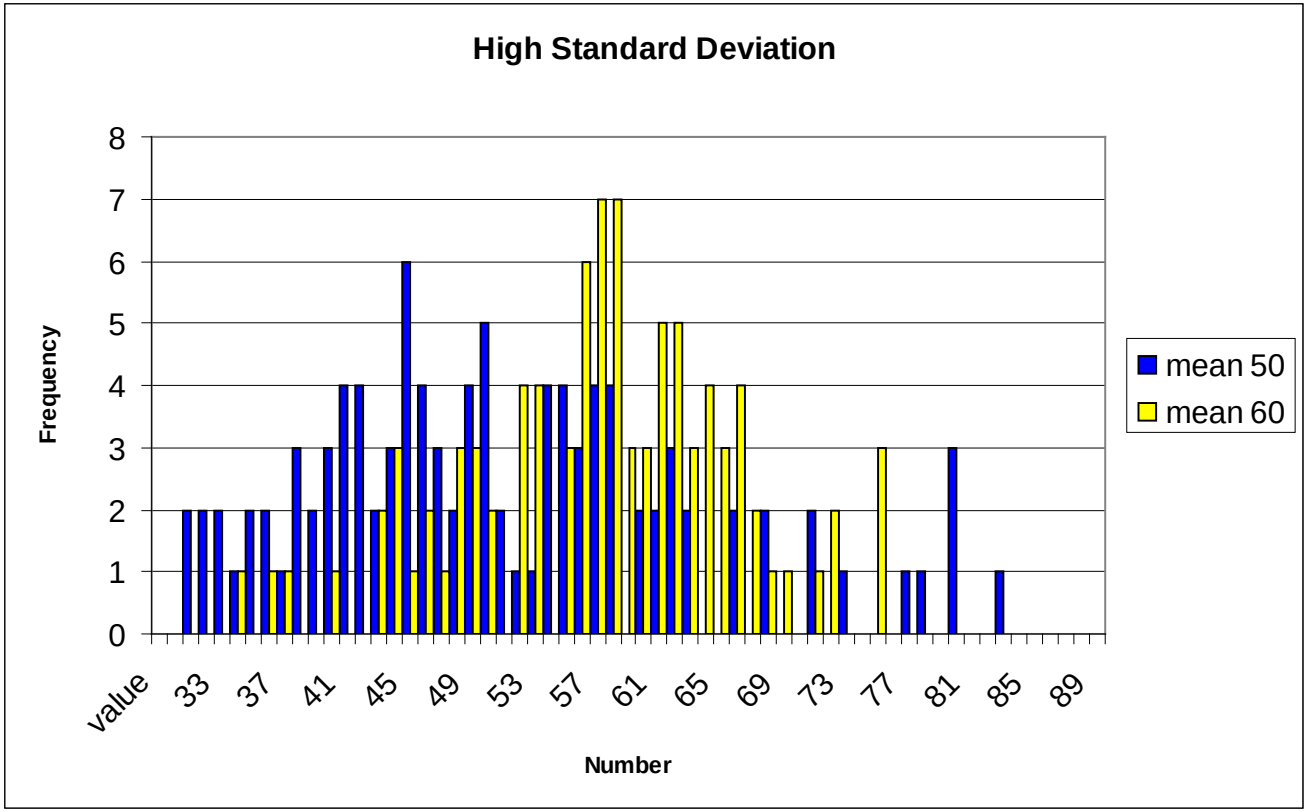
A tight conclusion



Less Precision



Can we conclude anything?



Two types of mistakes (Type I)

- Type I error: Reject the null hypothesis H_o when it is in fact true.
- The *significance level* (or *level*) of a test is the probability of a type I error
  - *A =P(Reject Ho|Ho is true)*
- Example:
  - The incentive schools score 0.15 SD higher than those in the control schools?
  - If I say they are different, how confident am I in the answer?
  - In other words, how confident can we be that this is a true difference as opposed to occurring by chance?
  - Common level of a: 0.05, 0.01, 0.1

Two types of mistakes (Type II)

- Type II error: Failing to reject H_o when it is in fact false.
  - The *power* of a test is one minus the probability of a type II error

    $$P(0) = P(\text{Reject } H_0 | \text{Effect size not zero})$$

  - How likely is my experiment to actually detect an effect if there actually is an effect?

Example: If I run 100 experiments, in how many of them will I be able to reject the hypothesis that incentive and control schools have the same test scores at the 5% level, if in fact they are different?

Power Calculations

- When planning an evaluation, with some preliminary research we can calculate the minimum sample we need to get to:
  - Test a pre-specified null hypothesis (e.g. treatment effect 0)
  - For a pre-specified significance level (e.g. 0.05)
  - Given a pre-specified effect size (e.g. 0.2 standard deviation of the outcomes of interest).
  - To achieve a given power

- A power of 80% tells us that, in 80% of the experiments of this sample size conducted in this population, if H_0 is in fact false (e.g. the treatment effect is not 0), we will be able to reject it.

- Power increases with:
  - Sample size (with the square root of sample size)
  - Minimum effect size (linearly with the effect size)
  - Design factors (stratification and baseline data help a lot)

Inputs for Power Calculations

What we need	Where we get it
Significance level	This is conventionally set at 5%
The mean and the variance of the outcome in the comparison group	From a small survey in the same or a similar population
The effect size that we want to detect	What is the smallest effect that should prompt a policy response? Rationale: If the effect is any smaller than this, then it is not interesting to distinguish it from zero

Picking an Effect Size

- What is the smallest effect that should justify the program to be adopted:
  - Cost of this program vs. the benefits it brings
  - Cost of this program vs. the alternative use of the money
- If the effect is smaller than that, it might as well be zero: we are not interested in proving that a very small effect is different from zero
- In contrast, any effect larger than that effect would justify adopting this program: we want to be able to distinguish it from zero
- Common effect sizes: 0.2 SD to 0.4 SD
  - Can detect smaller effects in some studies based on frequency of observations and intra-cluster correlations

The Design factors that influence power

- Clustered design
- Availability of a Baseline
- Availability of Control Variables, and Stratification.
- The type of hypothesis that is being tested
- Cost of Intervention

Clustered Design

Cluster randomized trials are experiments in which social units or clusters rather than individuals are randomly allocated to intervention groups

Examples:

PROGRESA	Village
Flipcharts, Performance-Pay	School
Balsakhi	Class
Iron supplementation	Family

Reasons for cluster randomization

- Need to minimize or remove contamination
  - Example: In the deworming program, schools was chosen as the unit because worms are contagious
- Basic Feasibility considerations
  - Example: The PROGRESA program would not have been politically feasible if some families were introduced and not others.
  - Same with the performance-pay interventions
- Only natural choice
  - Example: Any education intervention that affect an entire classroom (e.g. flipcharts, teacher training).

Impact of Clustering

- The outcomes for all the individuals within a unit may be correlated
  - All students are exposed to the same classroom
  - All students share the same teacher, school head, etc
  - The program affects all students at the same time.
  - The subjects of a program interact with each other
- Basic ideas is that 2 data points drawn from the same cluster are more ‘similar’ than 2 data points drawn completely at random
  - Example of measuring average height across the world
- We call ρ the correlation between the units within the same cluster

Implications For Design and Analysis

- Analysis: The standard errors will need to be adjusted to take into account the fact that the observations within a cluster are correlated.
- Adjustment factor (design effect) for given total sample size, clusters of size m, intra-cluster correlation of ρ, the size of smallest effect we can detect increases by compared to a non-clustered design
- Design: We need to take clustering into account when planning sample size

Implications For Design and Analysis

- We now need to consider ρ when choosing a sample size (as well as the other effects)
- It is extremely important to randomize an adequate number of groups
- Often the number of individuals within groups matter less than the number of groups

Availability of a Baseline

- Uses of a baseline
- Can check for balance between control and treatment groups before the treatment
  - Not crucial if randomization is correct but good to have!
- Reduce the sample size needed. Why?
  - Correlation between individual outcomes in pre and post program periods
  - The stronger the correlation, the bigger the gain.
  - Very big gains in power for very persistent outcomes such as tests scores
  - Explain intuition
- In general, very good to have a baseline
  - Typically the evaluation cost go up and the intervention cost go down

Stratified Samples

- Stratification will reduce the sample size needed to achieve a given power (you saw this in the Balsakhi exercise and also in the performance pay experiment).
- The reason is that it will reduce the variance of the outcome of interest in each strata (and hence increase the standardized effect size for any given effect size) and reduce the correlation of units within clusters
- Example: if you randomize within school and grade which class is treated and which class is control:
  - The variance of test score goes down because age is controlled for
  - The within cluster correlation goes down because the "common principal effect" disappears.
- Common stratification variables:
  - Baseline values of the outcomes when possible
  - We expect the treatment to vary in different subgroups

The Hypothesis that is being tested

- Are you interested in the difference between two treatments as well as the difference between treatment and control?
- Are you interested in the interaction between the treatments?
- Are you interested in testing whether the effect is different in different subpopulations?
- Does your design involve only partial compliance? (e.g. encouragement design?)
- Do you want to estimate elasticities?

The Cost of the Intervention

- Why does this matter?
- A data point is worth the same for precision regardless of whether it comes from the control or treatment groups (if default allocation is equal size)
- So a control data point is normally cheaper
- Indonesia example

Conclusions

- Power calculations involve some guess work.
- They also involve some pilot testing before the proper experiment begins
- They can tell you:
  - How many treatments to have
  - How to trade off more clusters vs. more observations per cluster
  - Whether the evaluation is feasible or not
- OD (Optimal Design) is a simple tool that you can use to calculate power and sample sizes – we will include practical exercises in the problem sets