Randomization as a Tool for Development Economists

Esther Duflo

Sendhil Mullainathan

BREAD-BIRS Summer school

Randomization as one solution

- Suppose you could do a Randomized evaluation of the microcredit program, how would you randomize?

Planning randomized evaluations: Different ways of randomizing

- Experiment: WISE iron supplementation study (Thomas et al.) 17,000 individuals. Randomization at the household level into treatment or placebo group.
- Pilot before launching large scale program: PROGRESA/Oportunidades pilot in 506 communities of a program that was planned to be expanded to 50,00 potential beneficiaries. Many researchers involved in analysis.

Types of Designs (2)

- Oversubscription schemes: Voucher lotteries. Example: Columbia voucher program (PACES). Angrist et al. Program is oversubscribed even among eligible students (poor, etc…). Government randomized among applicants (for fairness).
- Randomize phase in: Miguel and Kremer deworming study. 3 phases: 25 schools in 98, 25 schools in 99, 25 school sin 2000.

Types of design (3)

- Randomize at the class level within a school: Balsakhi Study (Banerjee et al.) Every school is treated. Works only if there is no externalities within school.

Types of designs (4)

- Uncertainty about different program designs require experimentation:
  - Seed in the Philippines (Ashraf, Karlan, Yin). Would clients be interested in commitment savings product?
  - Fertilizer adoption in Kenya (Duflo, Kremer, Robinson): what programs are the most likely to lead to increases in fertilizer adoption.
  - Take up of loan in South Africa (Bertrand, Mullainathan and others): what convinces people to take loans?

Types of designs (5)

- Encouragement designs: it is often possible to encourage take up of existing program (by advertising/helping some individuals to take advantage). Ex: $20 incentives to attend information session; help people to access existing pension program
- Randomize default option.
- What do those designs have in common?

Analysis with imperfect compliance:

- Denote Z (0,1) the initial randomization, and T (0,1) the treatment status.
- Z is randomly assigned, T is not unless T=Z.
- Comparison of outcomes for Z=1, Z=0 will give Intention to treat, which may or may not be the policy parameter of interest.
- You can use the randomization like an instrument, and just like IV, you get the treatment effect on the compliers.
- Simplest case is Wald estimate

$$E[Y|Z=1]- E[Y|Z=0]/(E[T|Z=1]- E[T|Z=0])$$

- Randomization often have much larger first stage than natural experiment, which increases their external validity.

Implementing and analyzing randomized evaluations

- Power issues:
  - Definition of power
  - Factors in power calculations:
    - Clustering
    - Stratification
    - Baseline/other control variables
    - Repeated measurement.

Hypothesis testing

Often we are interested in testing the hypothesis that the effect size is equal to zero:

We want to test:

$$H_o : \text{Effect size} = 0$$

Against:

$$H_a : \text{Effect size} \neq 0$$

(other possible alternatives: $H_a>0$, $H_a<0$).

Two types of mistakes

- 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 $= P(\text{Reject } H_o | H_o \text{ is true})$

  Example Hb in treatment group is 13.25 in wise experiment treatment group, and 13.12 in the control group. Are they the same?

  If I say no, how ~~confident am I in the answer~~?

Common level of α: 0.05, 0.01, 0.1.

Two types of mistakes

- 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

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

Example: If I run 100 experiments, in how many of them will I be able to reject the hypothesis that treatment and control have the same Hb levels at the 5% level? Or: how likely is my experiment to fail to detect an effect when there is one?

Calculating Power

- When planning an evaluation, with some preliminary research we can calculate the minimum sample we need to get to:
  - Test a pre-specified hypothesis (e.g. treatment effect is 0)
  - For a pre-specified 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_o is in fact false (e.g. the treatment effect is not zero), we will be able to reject it.

- The larger ~~the sample, the larger the power~~.

Common Power used: 80%, 90%

Ingredients for a power calculation in a simple study

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, or from pre-existing data
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

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.
- (Encouragement design or otherwise partial first stage).

And now for the problems

- Even though this is a randomized experiment and your sample is large enough, what problems might there be?

Threats to experiment validity

- Externalities:
  - Violates independence of potential outcomes with the instrument.
  - Leads to downward bias if externalities are positive, upwards otherwise.
- Examples:
  - Worms if randomization at the individual levels.
  - Control are expecting treatment in the future and change their behavior accordingly (Progresa?).

- Externalities: What to do:
  - With global externalities (e.g. prices), hard to deal with. Local externalities can be dealt with by randomizing at the appropriate level (e.g. schools).
  - Some designs make it possible to evaluate the presence/importance of externalities, e.g. fertilizer, worms.

Threats (2)

- Attrition
  - Non random attrition introduces a violation of independence assumption
  - Need to check whether attrition is different in T/C, characteristics of attritors are different.
  - Not sufficient, and control attrition is key. May be hard over long period etc…
  - There exist statistical methods to correct for attrition: parametric (Heckman etc…) or non-parametric (Manski bounds)

Threats (3)

- Experiment influences the data that is collected
  - E.G. incentives programs: may influence test score on the test that is being tested but not “learning” in a more general sense.
  - Evaluation itself may change behavior of treated or comparison groups.

Threats (3)

- With imperfect designs instrument may affected the people who do not get the treatment, violating independence and making IV inconsistent (even if ITT is not).
  - Externalities within treatment group
  - Instrument directly affects potential outcomes (e.g. vouchers: incentives to study may be larger)
  - Externalities across treatment and controls

Inference issues

- Accounting for design
  - Clustering
  - Stratification
- Multiple outcomes
  - Make outcome by outcome comparison between T and C difficult to interpret.
  - Possible adjustments: Bonferoni bounds, Effect size for a family (Katz, Kling, Liebman).

Inference issues (2)

- Sub-groups
  - Ex-post sub-groups analysis causes inference problems (back to specification searching).
  - Ex-ante, which subgroups to look for can be specified in design (grounded in a theory of why the intervention is likely to have a given effect), and ideally the experiment can be stratified by these subgroups
  - What if interesting things are found ex-post? We still believe they are useful. This is where replications are important (next replication study can be stratified by this characteristic).

Inference issues (3)

- Handling covariates
  - Same issue as subgroups/multiple outcomes
  - Need to specify them ex-ante
  - Can increase or decrease precision

From RE to Welfare Analysis

- Reduced forms versus total derivative
  - RE evaluate the total derivative with respect to the program (i.e. potential reaction of individuals to the program, provision of other inputs that are substitutes or complement)
  - RE typically evaluate a program that may do more than one thing: may not allow to separately identify parameters of interest.
    - Assess this by trying several programs (e.g. fertilizer: endorsement intervention).
    - Combine different evaluations where programs have different features (e.g. PROGRESA and successors).

- Partial and general equilibrium effects
  - Heckman critique: GE effects of a nationwide experiment may be very different from partial eq. effects from a small randomized evaluation.
  - E.g. vouchers.
  - It is one kind of externalities. It may be possible (though expensive) to randomize intervention at the level of the markets (e.g. indian villages).

- Can the results be generalized
  - It is a general problem with any evaluation.
  - There is some amount of replication, but not enough.
  - Need some guidance to determine what to replicate, what to vary (e.g. green textbooks vs red textbooks)
  - RE is as its most useful when it tests a well defined theory.
  - This is the case for RE that tests the educational production function, and for the new generations of RE that have been designed to test specific theories derived from literature (seed, fertilizer, “observing unobservables”).
  - Collecting information on chain of causality leading to the final outcomes, and processes on the how the program went

Randomized Evaluations and other evaluations

- We saw many ways to address the selection bias. We saw that all of them make assumptions, whose validity must be assessed as a function of the context
- There has been efforts to compare results from randomized and non-randomized methods: Lalonde, Heckman, Ichimura and Todd, others.
- These efforts have had mixed results, and there are not very many of them. More would be better…
- It is critical to do the non-randomized methods *before* getting the results of the experiment, otherwise publication bias/specification searching may induce too many false "accept".