Planning sample size for impact evaluations

David Evans, Banco Mundial

Basado en slides de Esther Duflo (J-PAL) y Jed Friedman (Banco Mundial)

Size of the sample for impact evaluations

- Pergunta geral

  ¿De que tamanho tem que ser a mostra para creivelmente perceber dado tamanho de impacto?

- ¿Que quer dizer «creívelmente» aqui?

  Tenho um nível de certeza que a diferença entre o grupo que recebeu o programa y o que não está devida ao programa

- A aleatorización tira os biases mas não tira o barulho: Funciona pela lei dos grandes números… ¿Que tão grande tem que ser o número?

¿Que tão grande?

- ¿2 pessoas selecionadas de uma maneira aleatoria?


- ¿10 personas?


- ¡Muitas personas! ¿Quantas são muitas?


Organização básica

- Ao final do experimento, comparamos o resultado de interés nos grupos de tratamento e de controle
- Nos interessa a diferença

Média do grupo de tratamento
– Médio do grupo controle
Tamanho do efeito

- Por exemplo

Renda a média de lares que recebem bolsa
– Renda a média de lares que não recebem bolsa
Tamanho do efeito

A estimação

Não temos suficiente dinheiro como para observar todos os lares senão uma mostra (nim temos que fazê-lo).

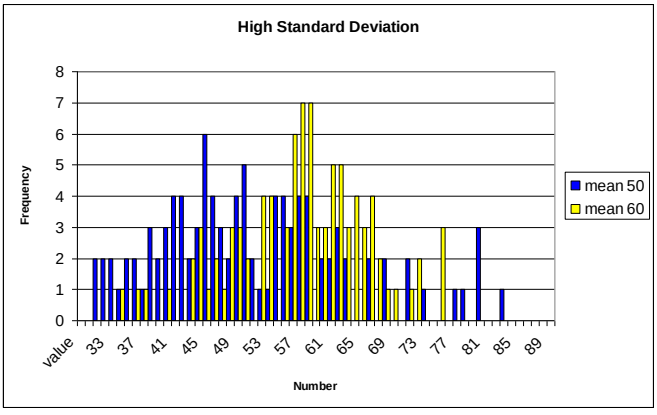
Em cada lar da mostra, há certo nível de renda. Pode estar mais perto ou mais longe da média de toda a população, como função de outros fatores que afeitam a renda.

Inferimos a renda média na população utilizando a média na mostra.

Se temos muito poucos lares, a médias estarão imprecisos. Se não vemos diferenças entre a média do grupo de tratamento e de controle, não sabemos se não há efeito ou se não há potencia de detectar o efeito.

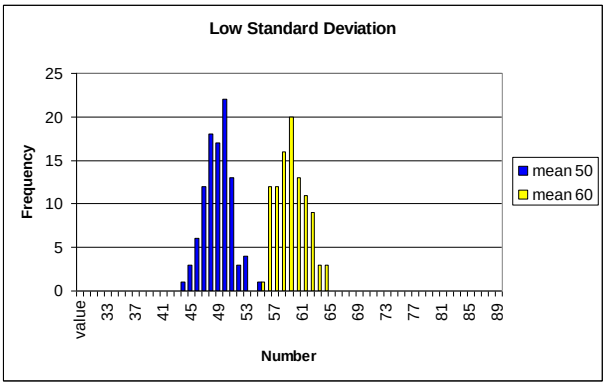
The variability that we measure in the result

If the results varies a lot within the treatment and the control group, it will be difficult to say whether it was the treatment that caused the difference in means



The variability that we measure in the result

If the result varies very little within the groups, it is easier to say that the treatment caused the difference



The standard error

- The standard error of the sample estimate captures the size of the sample and the variability of the result
  - ↑ with a small sample
  - ↑ with a high variable results
- A confidence interval of 95% for an effect tells us that, for 95% of the samples that we could draw from the same population, the estimated effect would fall in this interval.

Hypothesis Testing

Often we are interested in testing the hypothesis that the effect size is equal to zero (in other words, My program had no effect? I hope not!)

We want to test:

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

Against:

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

Two Types of Mistakes - 1

- First type of error : Conclude that there is an effect, when in fact there are no effect.

The level of your test is the *probability that you will falsely conclude that the program has an effect, when in fact it does not.*

So with a level of 5%, you can be 95% confident in the validity of your conclusion that the program had an effect

For policy purposes, we want to be very confident in the estimated impact: the level will be set fairly low.

Common level: 5%, 10%, 1%.

Relation with Confidence Intervals

- If zero does not belong to the 95% confidence interval of the effect size we measured, then we can be at least 95% sure that the effect size is not zero (thus there is an effect)
- So the rule of thumb is that if the effect size is more than twice the standard error, you can conclude with more than 95% certainty that the program had an effect

Two Types of Mistakes – 2

Second type of error: you think the program had no effect, when it fact it does have an effect.

- The Power of a test is the probability that I will be able to find a significant effect in my experiment if indeed there truly is an effect (higher power is better since I am more likely to find an effect)
- Power is a planning tool. It tells me how likely it is that I find a significant effect for a given sample size

Calculating Power

- When planning an evaluation, with some preliminary research we can calculate the minimum sample we need to get to:
  - Test a hypothesis: program effect was zero or not zero
  - For a pre-specified level (e.g. 5%)
  - Given a pre-specified effect size (what you think the program will do)
  - 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 there is indeed an effect in the population, we will be able to say in our sample that there is an effect with the level of confidence desired.
- 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 often conventionally set at 5%. The lower it is, the larger the sample size needed for a give power
The mean and the variability of the outcome in the comparison group	-From previous surveys conducted in similar settings - The larger the variability is, the larger the sample for a given power
The effect size that we want to detect	What is the smallest effect that should prompt a policy response? The smaller the effect size we want to detect, the larger a sample size we need for a given power

Picking an Effect Size

- What is the smallest effect that should justify the program to be adopted:
  - Cost of this program v the benefits it brings
  - Cost of this program v 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 danger: picking effect size that are too optimistic—the sample size may be set too low!

Standardized Effect Sizes

- How large an effect you can detect with a given sample depends on how variable the outcomes is.
  - Example: If all children have very similar learning level without a program, a very small impact will be easy to detect
- The standard deviation captures the variability in the outcome. The more variability, the higher the standard deviation is
- The Standardized effect size is the effect size divided by the standard deviation of the outcome
  - **d** = effect size/St.dev.
- Common effect sizes:

d=0.20 (small) d =0.40 (medium) d =0.50 (large)

The Design Factors that Influence Power

- The level of randomization
- Availability of a Baseline
- Availability of Control Variables, and Stratification.
- The type of hypothesis that is being tested.

Level of Randomization Clustered Design

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

Examples:

Conditional cash transfers	Villages
ITN distribution	Health clinics
IPT	Schools
Iron supplementation	Family

Reason for Adopting 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.
- 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 villagers are exposed to the same weather
  - All patients share a common health practitioner
  - All students share a schoolmaster
  - The program affect all students at the same time.
  - The member of a village interact with each other
- The sample size needs to be adjusted for this correlation
- The more correlation between the outcomes, the greater the need to expand the sample

Example of the effect of clustering

Number of classes, for power of 0.80

Intra-Class Correlation	Students in each class			
	10	50	100	200
0.00	23	7	5	4
0.02	25	10	8	8
0.05	30	16	15	13
0.10	40	25	23	22

Implications

- It is extremely important to randomize an adequate number of groups
- The number of individual within groups matter less than the number of groups
- The “law of large number” applies only when the number of groups that are randomized increase
- You CANNOT randomize at the level of the district, with one treated district and one control district! [Even with 2,000 students per district]

Design factors that influence power

- Level of the randomization
- Availability of a baseline
- Availability of control variables and stratification variables
- The kind of hypothesis you want to test

Availability of a Baseline

- A baseline has three main uses:
  - Can check whether control and treatment group were the same or different before the treatment
  - Reduce the sample size needed, but requires that you do a survey before starting the intervention: typically the evaluation cost go up and the intervention cost go down
  - Can be used to stratify and form subgroups
- To compute power with a baseline:
  - You need to know the correlation between two subsequent measurement of the outcome (for example: consumption measured in two years).
  - The stronger the correlation, the bigger the gain.
  - Very big gains for very persistent outcomes such as Labor Force Participation;

Los factores del diseño que influyen la potencia

- El nivel de la aleatorización
- La disponilidad de una línea de base (encuesta inicial)
- La disponilidad de variables control y de estratificación
- El tipo de hipótesis que se quiere poner a prueba

Control Variables

- If we have additional relevant variables (e.g. school size, student characteristics, etc.) we can also control for them
- What matters now for power is the variation that remains after controlling for those variables
- If the control variables explain a large part of the variance, the precision will increase and the sample size requirement decreases.
- Warning: control variables must only include variables that are not INFLUENCED by the treatment: variables that have been collected BEFORE the intervention.

Stratified Samples

- Stratification: create BLOCKS by value of the control variables and randomize within each block
- Stratification ensure that treatment and control groups are balanced in terms of these control variables.
- This reduces variance for two reasons:
  - it will reduce the variance of the outcome of interest in each strata
  - the correlation of units within clusters.
- Example: if you stratify by district for an agricultural prograr
  - Agroclimatic and associated epidemiologic factors are controlled for
  - The “common district government effect” disappears.

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.

The Hypothesis that is being Tested

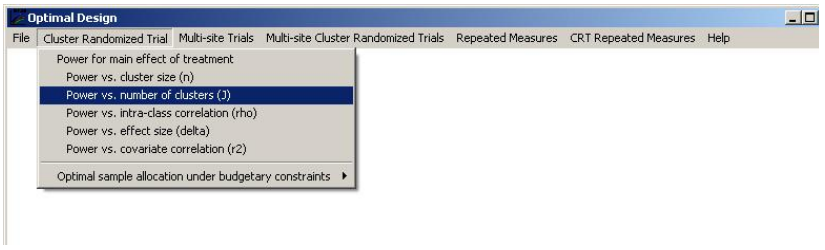
- Are you interested in the difference between two treatments as well as the difference between treatment and control?
- Are you interested in the interaction between the treatments?
- Are you interested in testing whether the effect is different in different subpopulations?
- Does your design involve only partial compliance?

To remember!

- The number of groups is much more important than the number of individuals
  - Schools vs students, villages vs homes
- Two types of errors
  - Type I: You think there is an effect when there is not → level
  - Type II: You think there is no effect then there is one → power
- Avoiding errors requires a sufficient sample → the power calculation

Power Calculations Using the OD Software

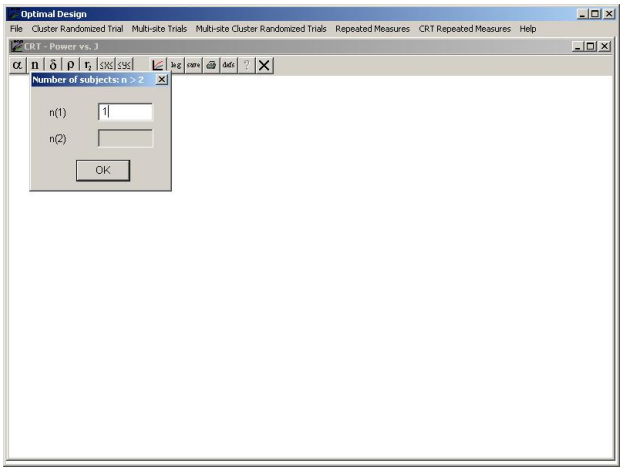
- Choose “Power v. number of clusters” in the menu “clustered randomized trials”



http://sitemaker.umich.edu/group-based/optimal_design_software

Cluster Size

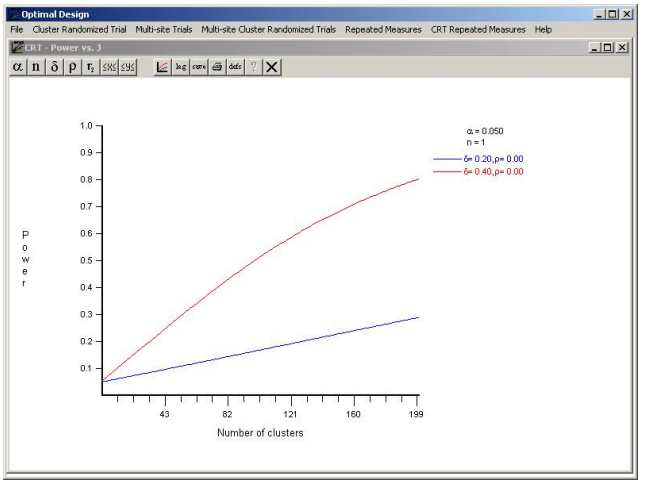
- Choose cluster size



Choose Significance Level, Treatment Effect, and Correlation

- Pick **a : level**
  - Normally you pick 0.05
- Pick **d :**
  - Can experiment with 0.20
- Pick the intra class correlation (rho)
- You obtain the resulting graph showing power as a function of sample size

Power and Sample Size



To remember!

- The number of groups is much more important than the number of individuals
  - Schools vs students, villages vs homes
- Two types of errors
  - Type I: You think there is an effect when there is not → level
  - Type II: You think there is no effect then there is one → power
- Avoiding errors requires a sufficient sample → the power calculation

Conclusions: Power Calculation in Practice

- Power calculations involve some guess work.
- At times we do not have the right information to conduct it very properly
- However, it is important to spend effort on them:
  - Avoid launching studies that will have no power at all: waste of time and money
  - Devote the appropriate resources to the studies that you decide to conduct (and not too much).