Planning Sample Size for Randomized Evaluations

Deon Filmer
Development Research Group
The World Bank

REGIONAL IMPACT EVALUATION WORKSHOP
Evaluating the Impact of Development Programs:
Turning Promises into Evidence

Seoul, Korea
December 6-10, 2010

Why sample?

- Can't do a survey of entire population because of …
  - Cost
  - Management of survey implementation

- In randomized evaluations we will always be comparing:
  - Mean in **treatment** group
  - Mean in **control** group

  → So we will need to estimate the group means using the averages in the samples

Why a **random** sample?

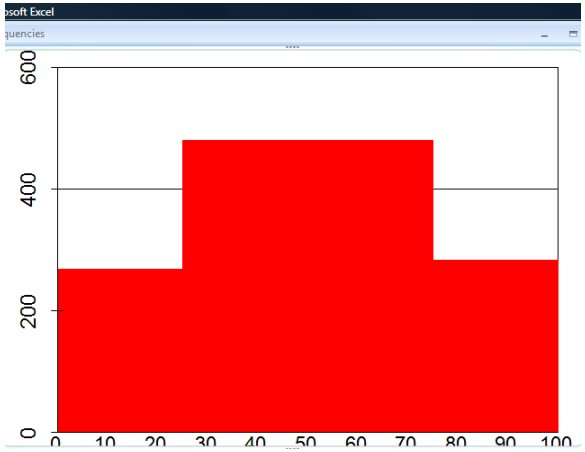
- Want the survey to be an accurate representation of the population as a whole
  - Randomness ensures that the sample reflects the population:
    - Randomization removes **bias** (that the estimate of the mean will be systematically wrong)
- How could I use a sample to estimate the percentage of women in the room?
  - Random/non-random

What is a good **sample size**?

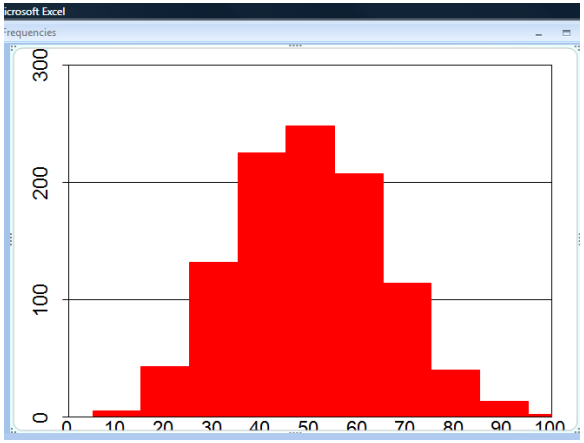
- Use a random sample to estimate percentage of women in the room.
- Different samples can give different results.
- So:
  - **What is the confidence we have in an estimate, given the sample size?**
- [Excel simulation #1]

Alternative sample sizes: increasing the confidence we have in the estimate

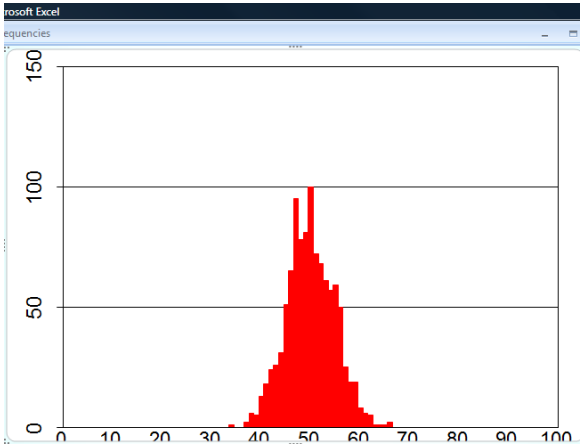
Sample size = 2



Sample size = 10



Sample size = 100



Why does this matter?

- When the confidence in estimates is high, this allows us to distinguish between two estimates (e.g. treatment and control).
- Remember: this is what we do in impact evaluation**: compare one estimate to another**
- [Excel simulation #2]

Why does this matter?

- We want to be confident that we won't mistakenly find an impact when there isn't one
- We want the **power** to be able to find an impact when there is one.

Level of a test

- The **level** of a test is
  - the probability that I will conclude that the program has an impact
  - when there **actually was no impact.**

- Common levels used: 5%, 10%, 1%

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

Power

- The **power** of a test is
  - the probability that I will be able to **actually find** a significant effect in my experiment
  - if there is **indeed truly** an effect

→ Higher power is better since I am more likely to report a true effect

- Power is a planning tool for study design. It tells me how likely it is that I find a significant effect for a given sample

Ingredients for a Power Calculation

What we need	Where we get it
Significance level (You decide)	This is often conventionally set at 5%. The lower it is, the larger the sample size needed for a give power
The effect size that we want to detect (effect size = expected impact scaled by the variability of the outcome)	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
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 required sample for a given power

Bottom line on sample size and power

- Sadly, no single answer to "**how big a sample do I need**" … it depends
  - How confident do I want to be about
    - Mistakenly saying there is an impact when there really isn't one
    - Identifying an impact when there is one
  - How big is the expected impact
  - How variable is the value of the outcome to begin with
- Larger samples allow you to distinguish smaller impacts

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
Bed net distribution	Health clinics
IPT	Schools
Social support	Family

Reasons for Adopting Cluster Randomization

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

Impact of Clustering

- The outcomes for all the individuals within a cluster may be correlated
  - All villagers are exposed to the **same** weather
  - All patients share a **common** health practitioner
  - All students **share** a schoolmaster
  - The members of a village **interact** with each other
- The sample size needs to be adjusted for this within-cluster correlation
- The bigger the correlation between the outcomes, the bigger the adjustment
- Consider the extreme (Everyone in the village is "the same")

Implications of clustering

- It is extremely important to sample an adequate number of randomly selected **clusters**
- (So, in the extreme, you CANNOT randomize at the level of the district, with one treated district and one control district!!!!)

Bottom line on **clustering**

- To get the same **power** (i.e. ability to distinguish treatment from control groups), a sample that uses clustering needs to be **larger**
- The **number of clusters** often matters more than the number of units (e.g. individuals, households) within clusters.

Sampling and power

- Larger samples allow us to better detect impacts—increasing the **power** of the evaluation to identify smaller impacts.
- **Clustering** means that samples will need to be larger still—and that there are enough clusters included in the sample.

Thank you