

SAMPLING AND STATISTICAL POWER

Erich Battistin

University of Padua

Kinnon Scott

DECRCG, World Bank

AADAPT Workshop

April 13, 2009

Introduction

- What are we trying to do with impact evaluation?
- Determine if an intervention or treatment has had an effect and what that effect is
- Because we cannot have information on the same person/community/farm in two different states at one time (no parallel universes) need to draw on sampling theory-some but not all answers
- Start with randomization as benchmark (applies to other designs)

What are we trying to do?

We want to test the hypothesis that the effect size is equal to zero:

We want to test:

H_o : Effect size = 0

Against:

H_a : Effect size > 0

Can be done for different groups of individuals

Basic Setup

- Randomly assign subjects to separate groups, each of which is offered a different “treatment”
- After the experiment, we compare the outcome of interest in the treatment and the control group
- We are interested in the difference:

Effect = Mean in treatment - Mean in control

*Example: average voting rate in intervention villages vis-à-vis
average voting rate in control villages*

*Change in production among treatment farmers compared to change
in production of control group of farmers*

Why randomize?

- Eliminates systematic pre-existing group differences (interest, wealth, entrepreneurship)
- However, randomization may produce experimental groups that differ by chance- not biases but *random errors*

Bottom line: randomization removes bias, but it does not remove random noise in the data

Basic Setup cont.

- We do not observe the entire population, just a sample. *Example: we do not have data for all villages of the country, but just for a random sample of them in treatment and control areas*
- We estimate the mean outcome of interest by computing the average in the sample. *Example: we compute the average pregnancy rate for villages in the sample to estimate the mean pregnancy rate in the population*

Bottom line:

Estimated Effect = True Effect + Noise

Planning Sample Size for Randomized Evaluations

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



Measure with a certain degree of confidence the difference between participants and non-participants

Key ingredients: number of **units** (e.g. villages) randomized, number of individuals (e.g. households) within units, info on the outcome of interest and the expected size of the effect

Hypothesis Testing

“Ideal” property of any testing procedure:

- minimize disappointment , but
- allow for a minimum degree of error

→ Avoid two types of mistakes

Type I Error

- Conclude that there is an effect of treatment, when in fact there are no effect
- SIGNIFICANCE LEVEL → probability that you will falsely conclude that the program has an effect,

$$~~H_a : \text{Effect size} > 0~~$$

when in fact it does not.

- For policy need to be very confident in the answer you give: so set level fairly low. Common levels are: 5%, 10%, 1% (with 5% significance level can be 95% confident concluding that program had an effect.

Type II Error

Fail to reject that the program had no effect, when in fact it does have an effect

$$~~H_0 : \text{Effect size} = 0~~$$

The power of a test is the probability that will be able to find a significant effect of the treatment if indeed there truly is one

Higher power is better since you are more likely to have an effect to report –avoid disappointment--and key for policy

Practical Steps

- Set a pre-specified confidence level (5%)
- Set a range of pre-specified effect sizes (what you think the program will do). What is the smallest effect that should prompt a policy response? Aka minimum detectable effect
- Decide on a sample size to achieve a given power (80% or 90%).
- Intuitively, the larger the sample, the larger the power. Power is a planning tool: one minus the power is the probability to be disappointed
- Budget.....

Practical Steps -- “magic formulas”

Proposition I:

There exists at least one statistician in the world who has already put into a magic formula the optimal sample size required to address this problem

Proposition II:

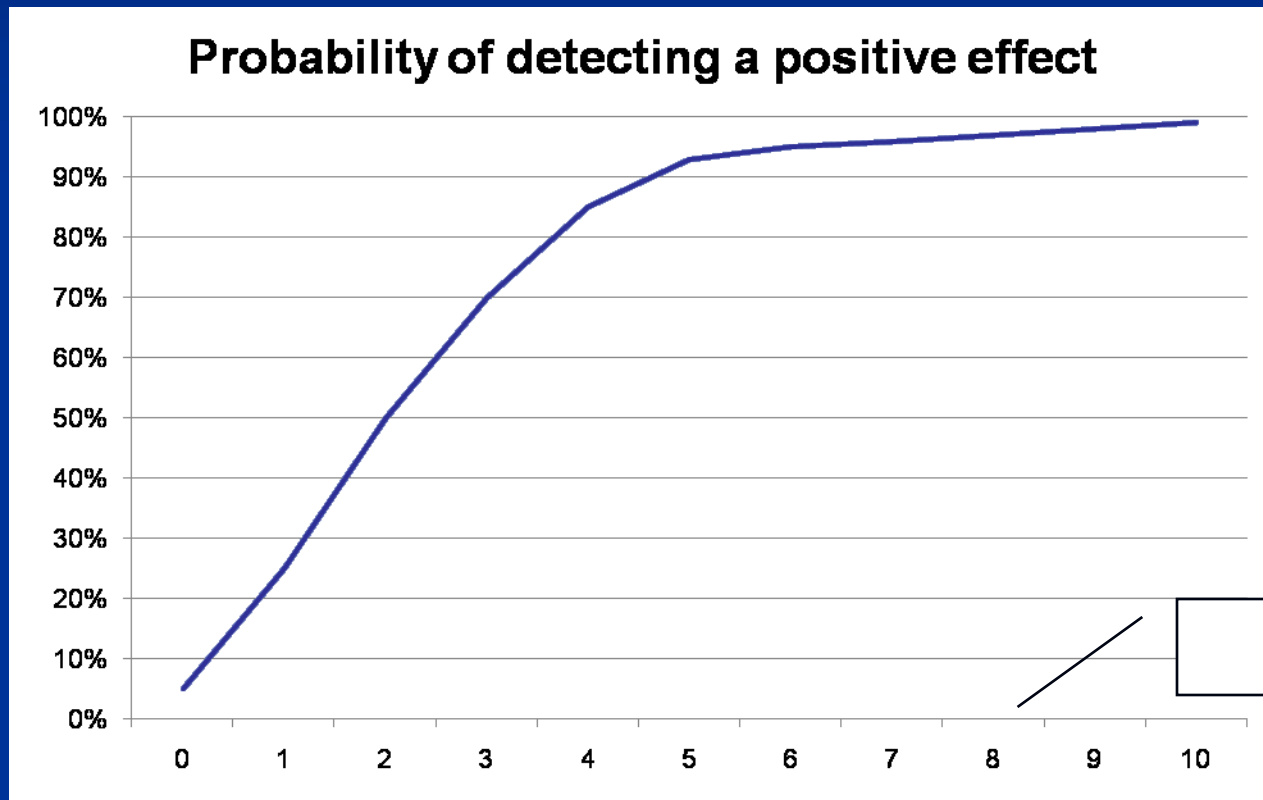
The rule has also been implemented for almost all computer software

Not difficult to do, and only requires simple calculations to understand the logic (really simple!)

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
- Common danger: picking effect size that are too optimistic—the sample size may be set too low

Hypothesis Testing, cont.



Sample size

General “rule”: the sample size required is a function of:

- **Significance level** (often set to 5%)
- **Minimum detectable effect** – you set this
- **Power to detect it** (often set to 80%)
- **Variance of the outcome of interest** before the intervention takes place (derived from baseline data)
- **Clustering**: effect of clustering (derived from baseline data)

The Design Factors that Influence Power

- The level of randomization (clustering)
- Availability of a Baseline
- Availability of Control Variables & Stratification

Level of Randomization

Clustered Design

- Cluster (or group) randomized trials are experiments in which social units or clusters rather than individuals are randomly allocated to the intervention group
- **Examples:** first randomize **villages**, and then observe outcome variables at the **household** level. Or, in an education program, randomize **schools** and then look at **students'** achievement.

Level of Randomization

Clustered Design (cont.)

- Cluster randomization provides unbiased estimates of intervention effects for the same reasons that individual randomization does
- However, the statistical power or precision of cluster randomization is less than that for individual randomization, and often by a lot!

Impact of Clustering

- The outcomes for all the individuals within a cluster may be correlated
 - All villagers are exposed to the same NGO
 - All patients share a common health practitioner
 - Inequality rates vary from village to village
 - The members of a village interact with each other
- The sample size needs to be adjusted for this correlation

The more correlation between the outcomes, the more we need to adjust the standard errors

Practical implications

- *It is extremely important to randomize an adequate number of clusters.*
- The general result is that the number of individuals within clusters matters less than the number of clusters
- Think that the “law of large number” applies only when the number of clusters that are randomized increases

Availability of a Baseline

A baseline has three main uses:

- Can get information on the outcome of interest before the intervention is implemented
- Can check whether control and treatment group were the same or different before the treatment (this may turn out very useful in non-experimental settings)
- Can be used to stratify and form subgroups

Control Variables

- To improve precision or to ensure that specific groups can be analyzed (gender, ethnicity, certain crops) one can stratify experimental sample members by some combination of their baseline characteristics, and then randomize within each stratum
- Factors used for stratifying in social research typically include
 - geographic location,
 - demographic characteristics,
 - past outcomes

Control Variables (cont.)

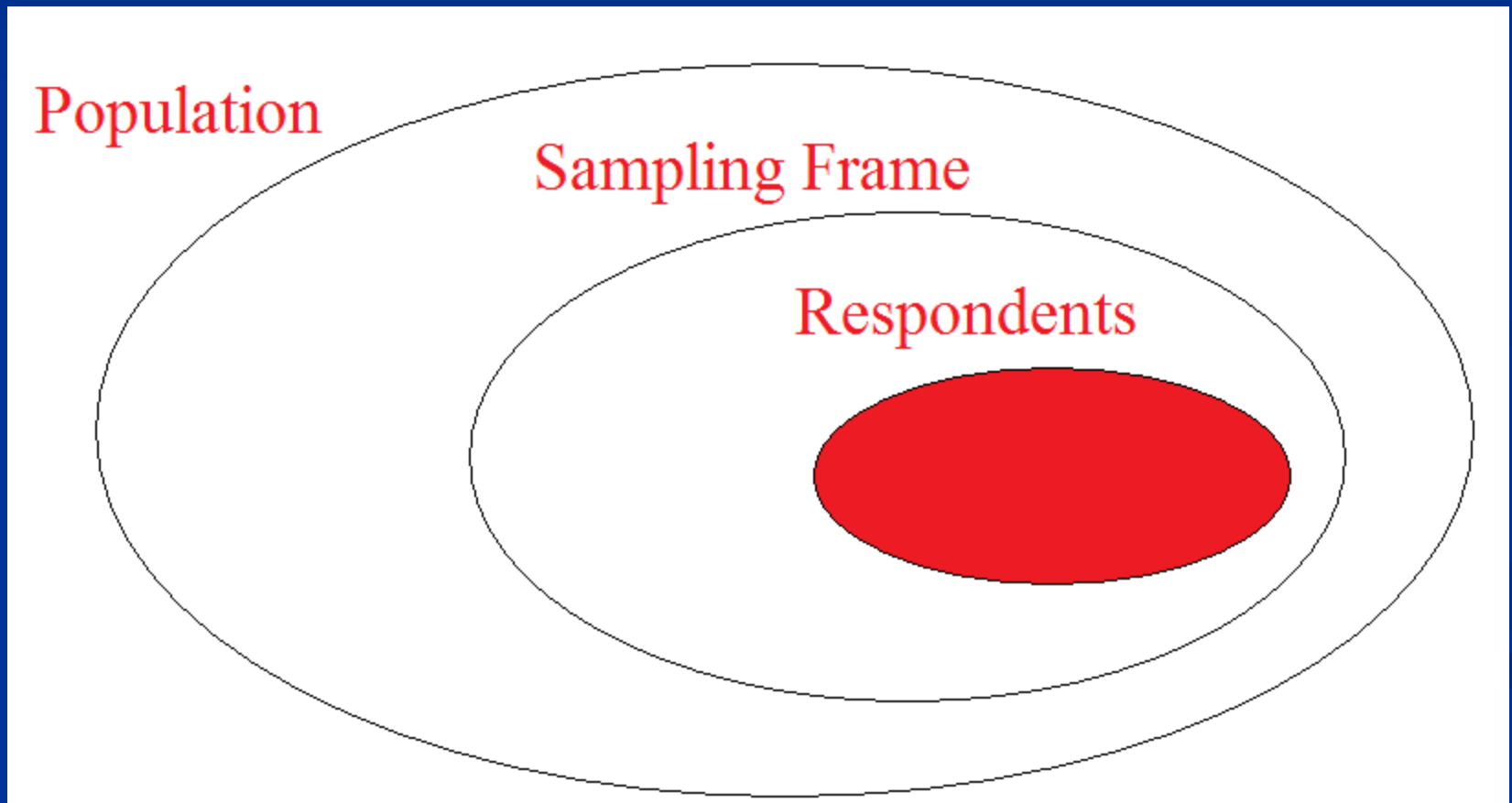
- If the control variables explain a large part of the variance, the precision will increase and the sample size requirement decreases. This reduces variance for two reasons:
 - reduces the variance of the outcome of interest in each stratum, and
 - the correlation of units within clusters
- **Warning: control variables must only include variables that are not INFLUENCED by the treatment, i.e. variables that have been collected BEFORE the intervention**

Control Variables (cont.)

- What matters now for power is the residual variation after controlling for those variables... so just replicate the steps described above within strata
- It may help stratifying along dimension that we know from previous studies are important for the effects of the programme. **Example:** we might expect to have differential effects by gender or age groups
- This may help understand “non-response” rates

Non-response

Graphically



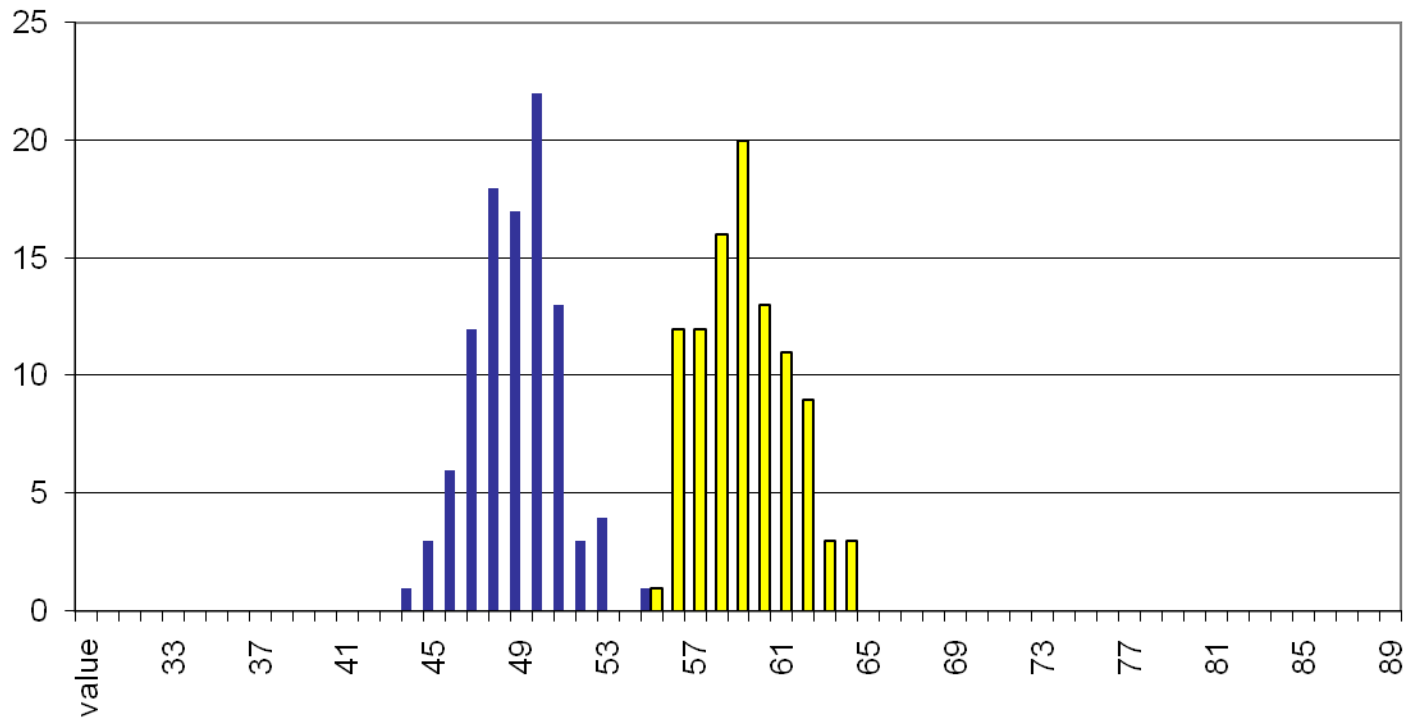
Summary

- Power calculations involve some guess work
- At times we do not have the right information to do it very well
- 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)
- Budget

Thank You
Merci
Obrigada

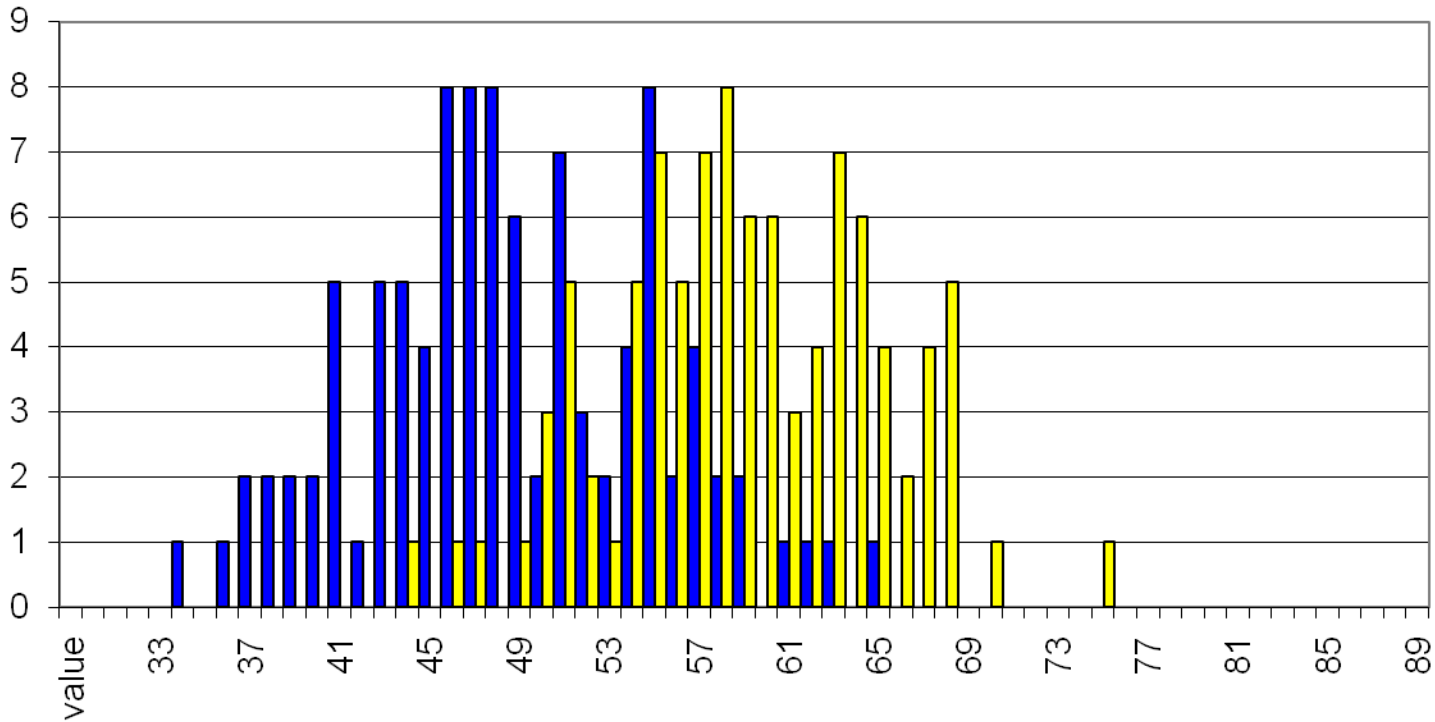
What do we mean by “noise”?

Precisely Estimated: Low Variance



What do we mean by “noise”?

Imprecisely Estimated: Large Variance



Relation with Confidence Intervals

- A 95% confidence interval for an effect size tells us that, for 95% of any samples that we could have drawn from the same population, the estimated effect would have fallen into this interval
- 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
- 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

Standardized Effect Sizes

(but this is a 2OP here)

- Sometimes impacts are measured as a standardized mean difference, for example when outcomes in different metrics must be combined or compared
- The standardized mean effect size equals the difference in mean outcomes for the treatment group and control group, divided by the standard deviation of outcomes across subjects within experimental groups