Sampling for Evaluation

Mattias Lundberg, World Bank
SIEF Regional Impact Evaluation Workshop
Accra, Ghana
May 2010

Adapted from slides by Esther Duflo and Jed Friedman

Outline

- **Why sample?**
- **Sample frames and sampling methods**
- **Hypothesis testing**
- **Sample size and power**
- **Issues affecting power**
- **Example using OD software**

Why use samples?

- Because we can't clone.
- Because we don't have enough money or time to study the entire population.
- Because we want to understand the impact of a program among a real population.

What do we want to know?

- Experiments are expensive. Do we have a chance of finding an impact?
- How big a sample do we need in order to be likely to be able to detect an impact?
- How do different experimental designs affect our chance to find an impact?

The sample frame

- Start with the population of interest (the group you want to learn about or to which you want to apply general lessons):
  - General population
  - Poor households
  - Young women
- Define the sub-population that you can in principle study (those you have access to)
  - Young women in region *x*, at time *t*….
- Make a list of those that you can actually contact
  - Names and contact information

This is your sample frame →

Note that this sample may not exactly represent the population of interest

The sample frame

- ***Recent*** ready-made lists provide the ideal sample frame:
  - Census
  - Telephone book
  - Voter rolls

But what if you don't have one of these?

- Conduct an initial enumeration – go into your study areas and make your own lists.
- Or don't – and bear the consequences.

The sample frame

So what are the consequences? In principle:

- with a sampling frame, you can draw general lessons for the wider population;
- without the sample frame, you can't generalize beyond the sample population.
- *NB: we sometimes do anyways.*

For example: a training program advertises widely and holds an information session to recruit participants. The study population comprises those who choose to come to the session.

We *should* only draw lessons for "those who show up to information sessions;" but we sometimes draw wider conclusions.

Sampling

Selecting a sample from the sample frame:

- **Probability samples (with known probabilities of selection)**
  - ***Simple Random* samples (choose randomly from entire list).**
  - ***Systematic* samples (choose first one randomly, then every n^{th} person from the list).**
  - ***Stratified* samples (divide the population into unique groups *i* such that $\sum_{1}^{I} n_i = N$, and sample within each group).**
  - ***Cluster* samples (randomly sample groups [census tracts, schools] first, then measure all units within the selected groups).**

Sampling

Selecting a sample from the sample frame:

- Non-probability samples
  - *Convenience* (eg clients at a health facility on one day)
  - *Purposive* (eg hard-to-reach populations, modal ['most likely'] samples, quota samples, heterogeneity samples, snowball samples)

If you depart from probability sampling, think about the impact of your non-random selection methods for your conclusions and for the generalizability of your results.

Hypothesis testing

- We want to test the "null hypothesis" that the impact of an intervention is zero,

$$H_0: impact = 0;$$

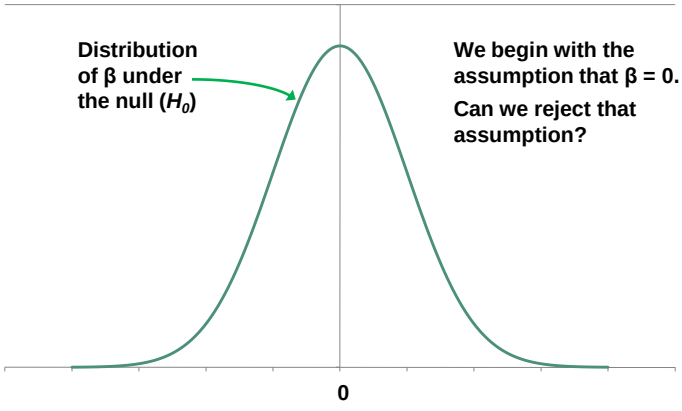
against the alternative hypothesis that the impact of the intervention is not zero,

$$H_a: impact \neq 0.$$

- We do this by comparing mean outcomes among a treated group to mean outcomes among a non-treated group.

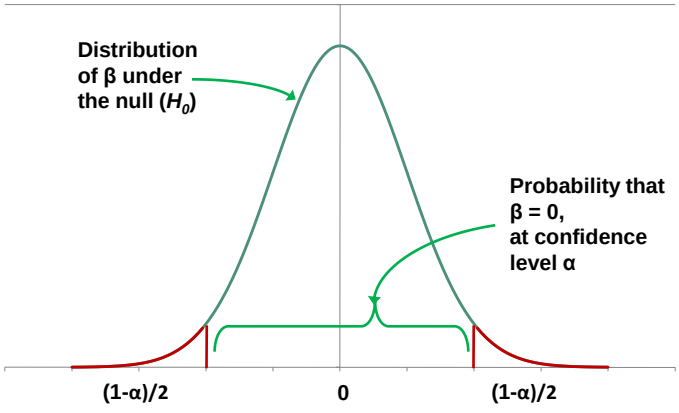
Hypothesis testing

- The expected distribution (the null hypothesis)



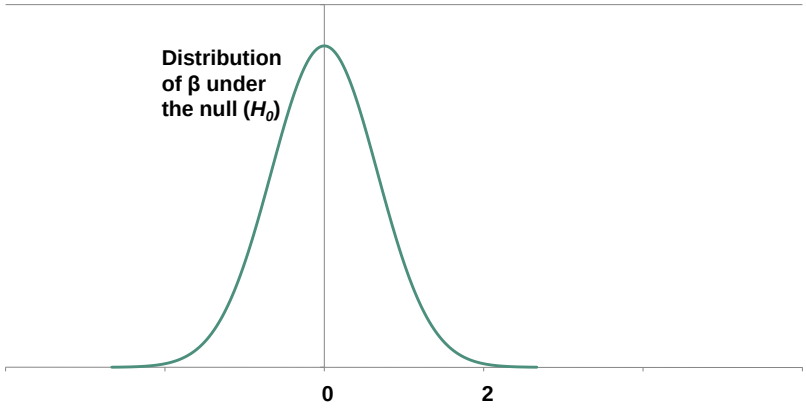
Hypothesis testing

- The expected distribution (the null hypothesis)



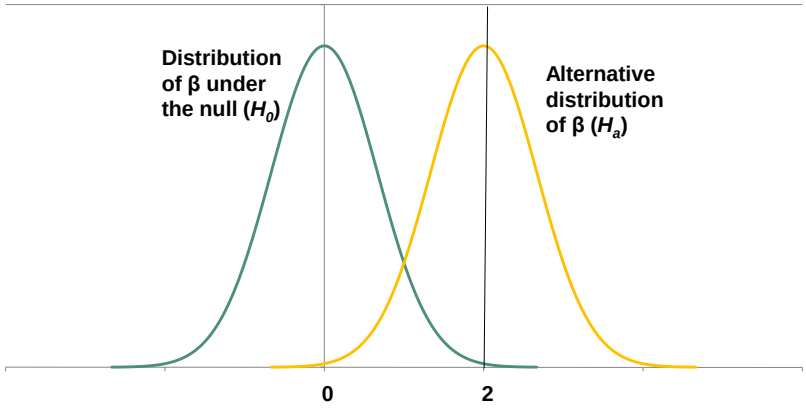
Hypothesis testing

- The expected and the alternative distribution



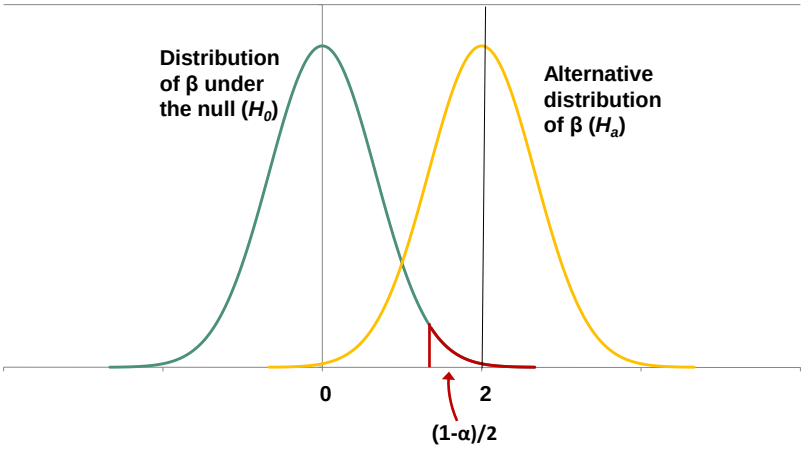
Hypothesis testing

- The expected and the alternative distribution



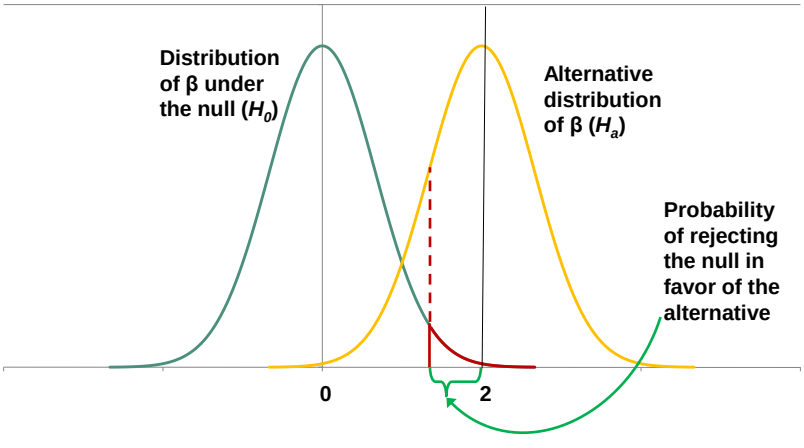
Hypothesis testing

- Are H_0 and H_a statistically significantly different?



Hypothesis testing

- Are H_0 and H_a statistically significantly different?



Confidence, power, and two types of mistakes

- First type of error: conclude that there is an effect, when in fact there is no effect (*false positive*).
- Second type of error: conclude that there is no effect, when in fact there is an effect (*false negative*).

		State of the world	
		H_0	H_a
Estimate	H_0	Correct acceptance	Incorrect acceptance (false negative, type-II, β)
	H_a	Incorrect rejection (false positive, type-I, α)	Correct rejection

Confidence, power, and two types of mistakes

- *Confidence* describes the test's ability to minimize type-I errors (false positives).
- *Power* describes the test's ability to minimize type-II errors (false negatives).
- The convention is to be much more concerned with type-I errors than type-II errors

  (ie, we're more willing to mistakenly say that something didn't work when it actually did, than to say that something worked when it actually didn't).
- We usually want *confidence* to be at least 90 or 95 percent, but will settle for *power* of 80 or 90 percent.

Calculating power

- What information do you need?
  - *Effect size* (minimum acceptable difference between means of treated and untreated groups, or expected difference between them)
  - *Sample variance* (standard deviations of means of treated and untreated estimates)
  - The acceptable *confidence level* (eg, 95%)
- With that information, you can estimate the sample size you need to achieve a certain power, or the power you will achieve given a certain sample size.

Calculating power

- Finding effect sizes and variance estimates:
  - Choose the smallest effect that would justify adoption of the program:
    - compare the cost of the program to the total value of its benefits.
  - Use evidence from previous comparable studies.
- Small effects will be harder to find if the outcome is highly variable in the population, or if it is imprecisely measured.

Why does size matter?

- Imagine we want to test the impact of a treatment ***T*** on an outcome ***Y***:

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

$\hat{\beta}$ is a measure of the average treatment effect: $\hat{\beta} = \bar{Y}_T - \bar{Y}_0$

- The variance of the estimated $\hat{\beta}$ is:

$$\text{var}\left(\hat{\beta}\right) = \frac{1}{P(1-P)}\frac{\sigma^2}{N}$$

where ***P*** is the proportion treated and ***N*** is the sample size.

Why does size matter?

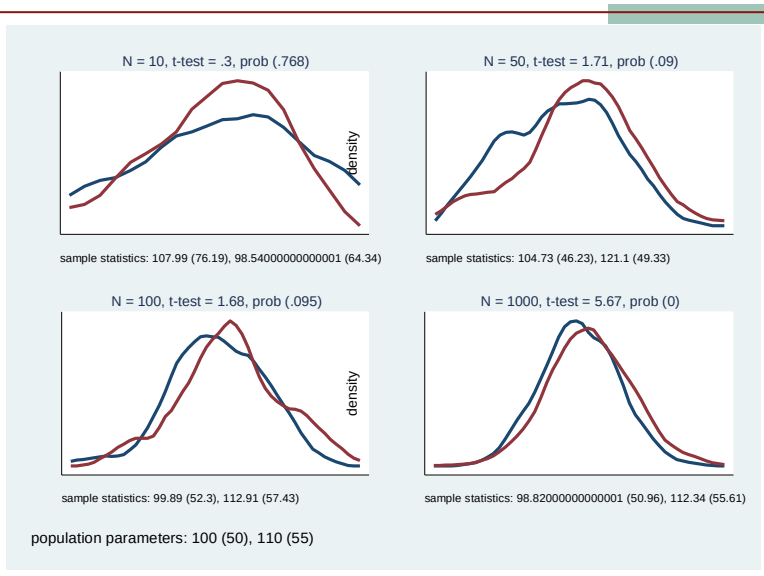
$$\text{var}\left(\hat{\beta}\right) = \frac{1}{P(1-P)}\frac{\sigma^2}{N}$$

- $\text{var}\left(\hat{\beta}\right)$ decreases as ***N*** increases, as σ^2 decreases, and as ***P*** → **0.5**.
- You're more likely to find a significant effect with a larger $\hat{\beta}$ and smaller $\text{var}\left(\hat{\beta}\right)$:

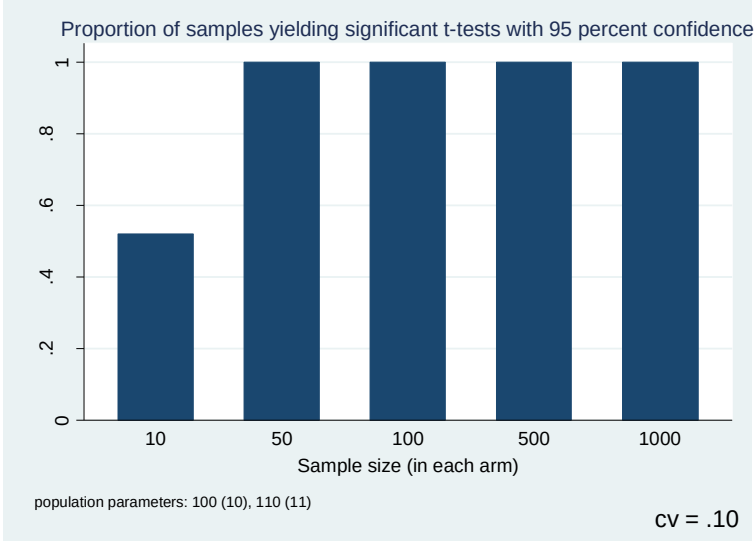
$$\frac{\hat{\beta}}{\sqrt{\text{var}\left(\hat{\beta}\right)}} \sim t_n$$

(the familiar *t*-statistic)

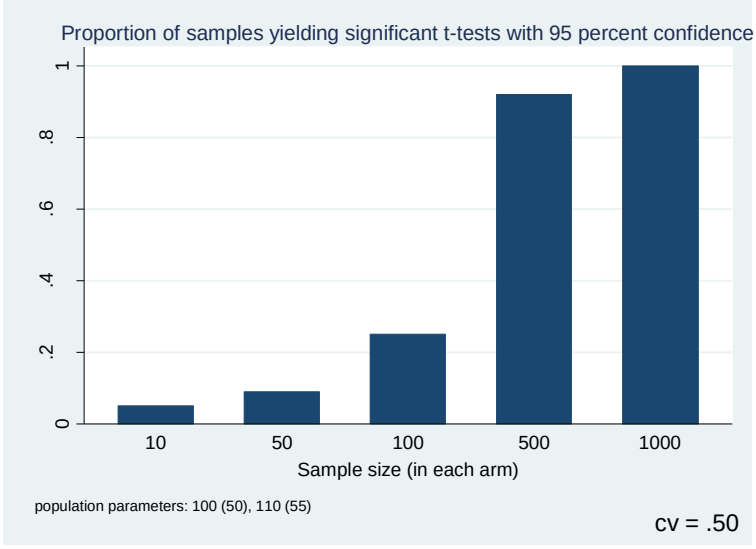
Power and sample size



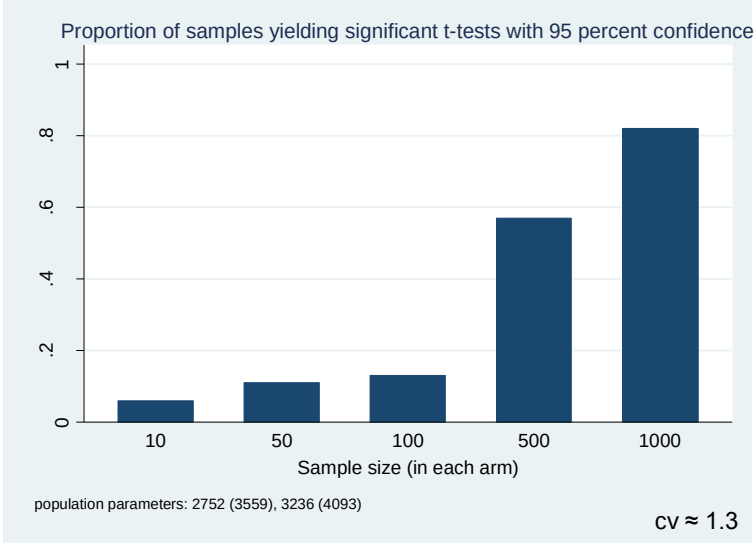
Power and variance
accurately measured outcomes



Power and variance
imprecisely measured outcomes



Power and variance
and using the Card (2007) parameters



Power and variance
and using the Card (2007) parameters

- The Card sample had 563 in the control group and 786 in the treatment group – a ratio of roughly 1.4.
- The mean outcomes (stddev) were 2752 (3559) in the control group, 3236 (4093) in the treatment group.
- They were lucky to have found a difference.
- In order to ensure measurement with 95 percent confidence, with 80 percent power, they should have had 826 in the control group and 1157 in the treatment group (that's a 50 percent increase).

Cluster Sampling

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

Examples:

Intervention	Cluster
Conditional cash transfers	Villages
Bed net distribution	Health clinics
Community management	Schools
Social support	Family

Why might you want to use cluster sampling?

- Spillovers, contagion or contamination
  - An intervention affects an entire group (e.g. teacher training affects an entire class, a community development program affects an entire village)
  - In one deworming program study, schools were chosen as the unit because worms are contagious
- Social or political considerations
  - The PROGRESA program would not have been politically feasible if some families in a village were introduced and not others.
  - A reintegration program for ex-combatants needed to reduce the likelihood of intra-village conflict.

Implications of clustering

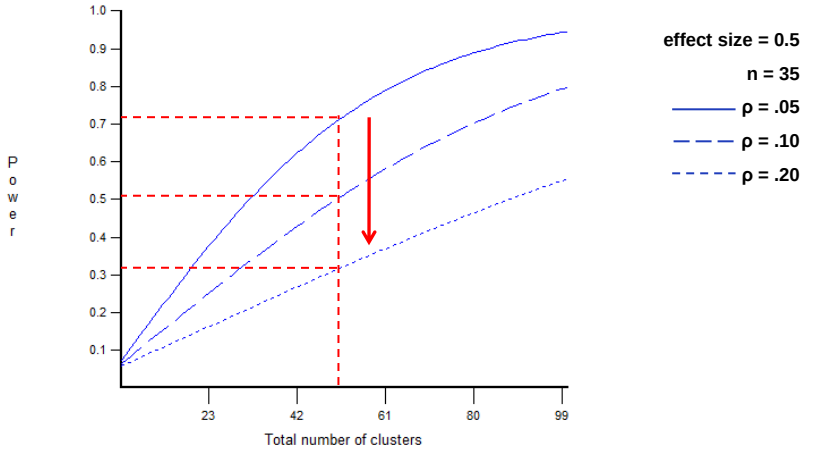
- Outcomes, responses to an intervention, for individuals within a group may be correlated:
  - All villagers are exposed to the same weather.
  - All patients share a common health practitioner.
  - All students share a schoolmaster.
  - Village residents interact with each other.
- The estimated standard errors must be adjusted to account for this correlation.
- The power of the sample falls as the intra-group correlation rises (because each individual provides less independent information).

Implications of clustering

- Make sure that you have a large enough sample.
- The primary sample is the number of units (people, families, villages, schools) at the level where the random assignment takes place. It is ***not*** the same as the number of people surveyed.
- For example, you ***cannot*** randomize at the level of the district, with a sample of one treated district and one control district, even if each district has a thousand people.
- The number of individuals within groups generally matters less than the number of groups (because of intra-cluster correlation).

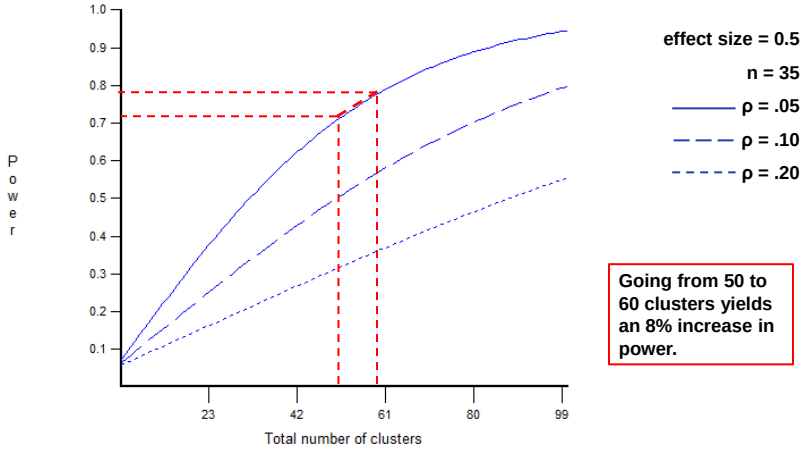
Intra-cluster correlation reduces power

- As intra-cluster correlation increases, each observation within each cluster provides less independent information.
- With 50 clusters, each with 35 observations:



Power can increase faster with the number of clusters

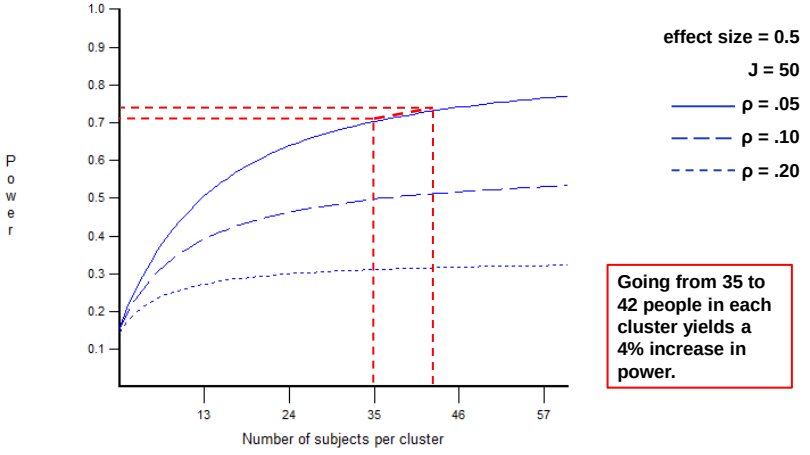
- You have 50 clusters with 35 people in each (1750 people), and you can afford to increase the sample by 20%, to 2100.
- Do you increase the size of each cluster or the number of clusters?



Going from 50 to 60 clusters yields an 8% increase in power.

than with the number of units within each cluster

- You have 50 clusters with 35 people in each (1750 people), and you can afford to increase the sample by 20%, to 2100.
- Do you increase the size of each cluster or the number of clusters?



Going from 35 to 42 people in each cluster yields a 4% increase in power.

Sampling with a baseline survey

- Advantages of a baseline:
  - Can check whether control and treatment group were the same before treatment.
  - Can be used to look at changes as well as levels.
  - Reduce the sample size needed, but requires that you do a survey before starting the intervention.
  - Can be used to stratify and form subgroups.
- To compute power with a baseline, you must know the correlation between subsequent measures of the outcome (for example, consumption measured in two years)

Other issues that affect sample size and selection

- Is there more than one treatment (more than one experimental arm)?
- Are you interested in differences between treated groups?
- Are you interested in interactions among treatments?
- Are you interested in testing whether the effect is different in different subpopulations?
- You must calculate power for *each of these comparisons*, not the intervention as a whole.

Other issues that affect sample size and selection

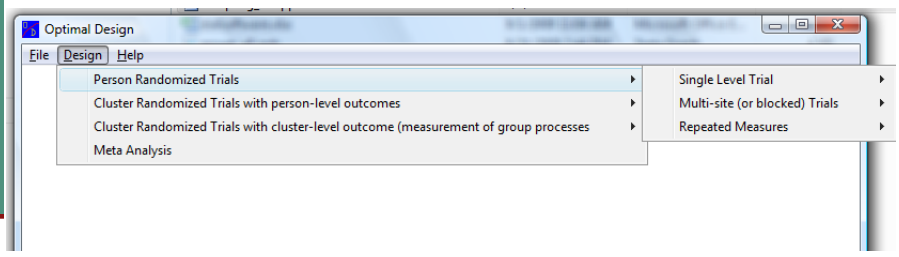
- Does your design involve only partial compliance? (e.g. encouragement design?)
- What will happen to the validity of your experiment if some people drop out?
- What if the attrition is not random?
- What if the intervention is not implemented according to plan?
- NB: power calculations are conditional and focus on one variable at a time; multivariate power calculations (involving joint distributions) are extremely difficult.

Software for power calculations

Proprietary	
Stata	http://www.stata.com/
SAS	http://www.sas.com/
SPSS	http://www.spss.com/
Splus	http://www.insightful.com/
Free	
PS: Power and Sample Size Calculation	http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/PowerSampleSize
Optimal Design Software	http://sitemaker.umich.edu/group-based/optimal_design_software
Web-based, interactive	
	http://statpages.org/
	http://www.dssresearch.com/toolkit/spcalc/power_a2.asp

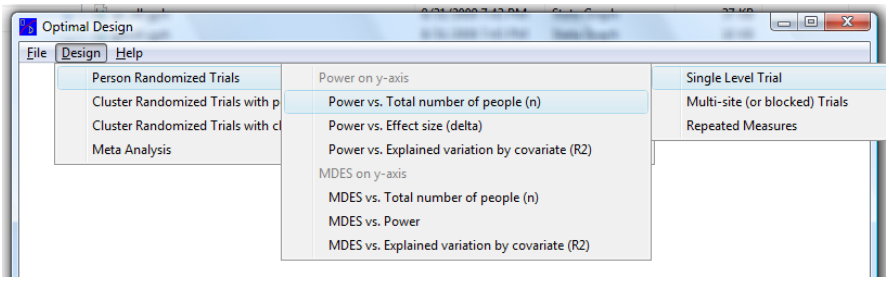
Example using OD software (1)

- To calculate sample size needed for a single person-level (unclustered) test:



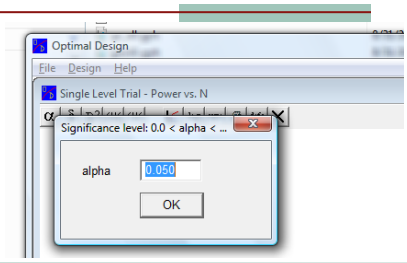
Example using OD software (1)

- OD can express the answer in terms of power or in terms of minimum detectable effect size.
- In this case, let's ask for power and sample size at a given effect size:

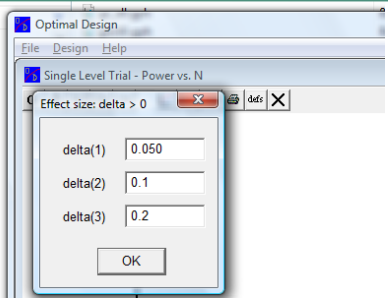


Example using OD software (1)

- First select desired significance level ***α*** (e.g. 5%)



- Then select minimum acceptable effect size ***δ*** (you can choose up to three)

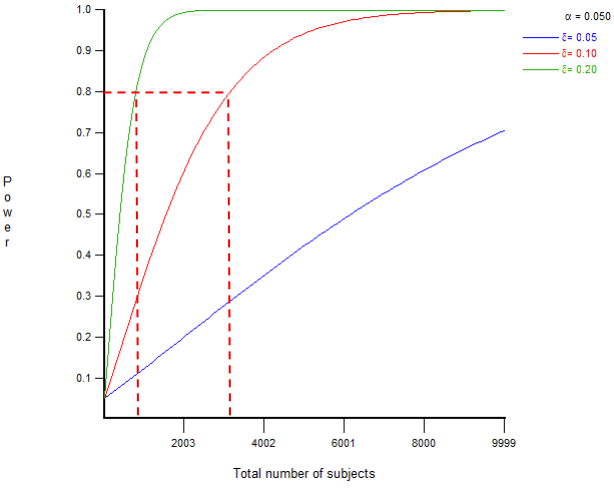


Example using OD software (1)

With an effect size of 0.2, 80% power is achieved with a sample of ~800.

With an effect size of 0.1, you need ~3100.

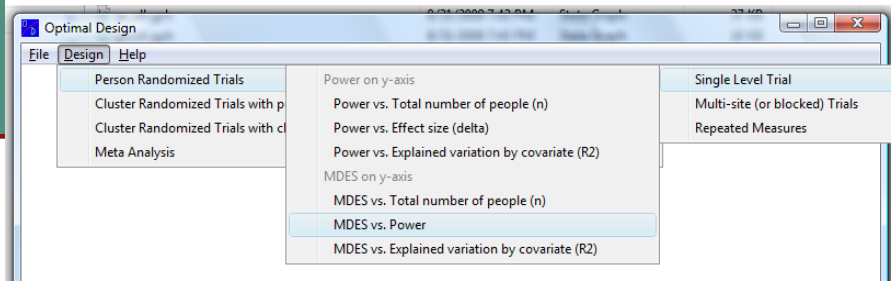
Smaller effects are harder to verify.



NB: X-axis scale changed from the default.

Example using OD software (2)

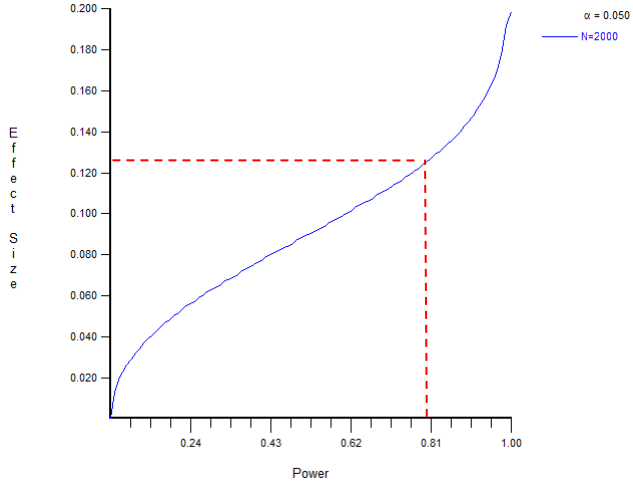
- Alternatively, if you know you can afford a sample of 2000, you can ask for the minimum detectable effect size given that sample:



Example using OD software (2)

With a sample of 2000, you can achieve 80% power at an effect size of ~0.125.

If you expect a smaller effect, you should ask for a bigger budget.



NB: Y-axis scale changed from the default.

Conclusions: Power Calculation in Practice

- Power calculations involve some guess work – we don't always have all the right information.
  - They can help you to avoid wasting time and money by launching studies that will have no power at all.
  - They can support arguments that you must devote sufficient resources to the studies that you decide to conduct.
- Anticipate; have back-up or contingency plans; know how you will deal with deviations from perfect sampling ex post. It will happen.

Resources on Sampling and Statistics for Evaluation

- UNDESA (2005) *Household Sample Surveys in Developing and Transition Countries*
  http://unstats.un.org/unsd/hhsurveys/pdf/Household_surveys.pdf
- Baker, J. (2000) *Evaluating the Impact of Development Projects on Poverty: A Handbook for Practitioners*
  http://go.worldbank.org/8E2ZTGBOI0
- Bamberger, M. (2006) *Conducting Quality Impact Evaluations under Budget, Time and Data Constraints*
  http://lnweb90.worldbank.org/oed/oeddoclib.nsf/DocUNIDViewForJavaSearch/757A5CC0BAE22558852571770059D89C/$file/conduct_qual_impact.pdf