Planning Sample Size for Randomized Evaluations

Jed Friedman, World Bank

SIEF Regional Impact Evaluation Workshop

Beijing, China

July 2009

Adapted from slides by Esther Duflo, J-PAL

Planning Sample Size for Randomized Evaluations

- General question:

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

- What does “Credibly” mean here?

  It means that I can be reasonably sure that the difference between the group that received the program and the group that did not is due to the program

- Randomization removes bias, but it does not remove noise: it works because of the law of large numbers… how large much large be?

Basic set up

- At the end of an experiment, we will compare the outcome of interest in the treatment and the comparison groups.
- We are interested in the difference:

  Mean in treatment - Mean in control

  = Effect size
- For example: mean of the number of adopted bed nets in villages with free distribution v. mean of the number of adopted bed nets in villages with cost recovery

Estimation

But we do not observe the entire population, just a sample

In each village of the sample, there is a given number of bed nets. It is more or less close to the actual mean in the total population, as a function of all the other factors that affect the number of bed nets

We estimate the mean by computing the average in the sample

If we have very few villages, the averages are imprecise. When we see a difference in sample averages, we do not know whether it comes from the effect of the treatment or from something else

Estimation

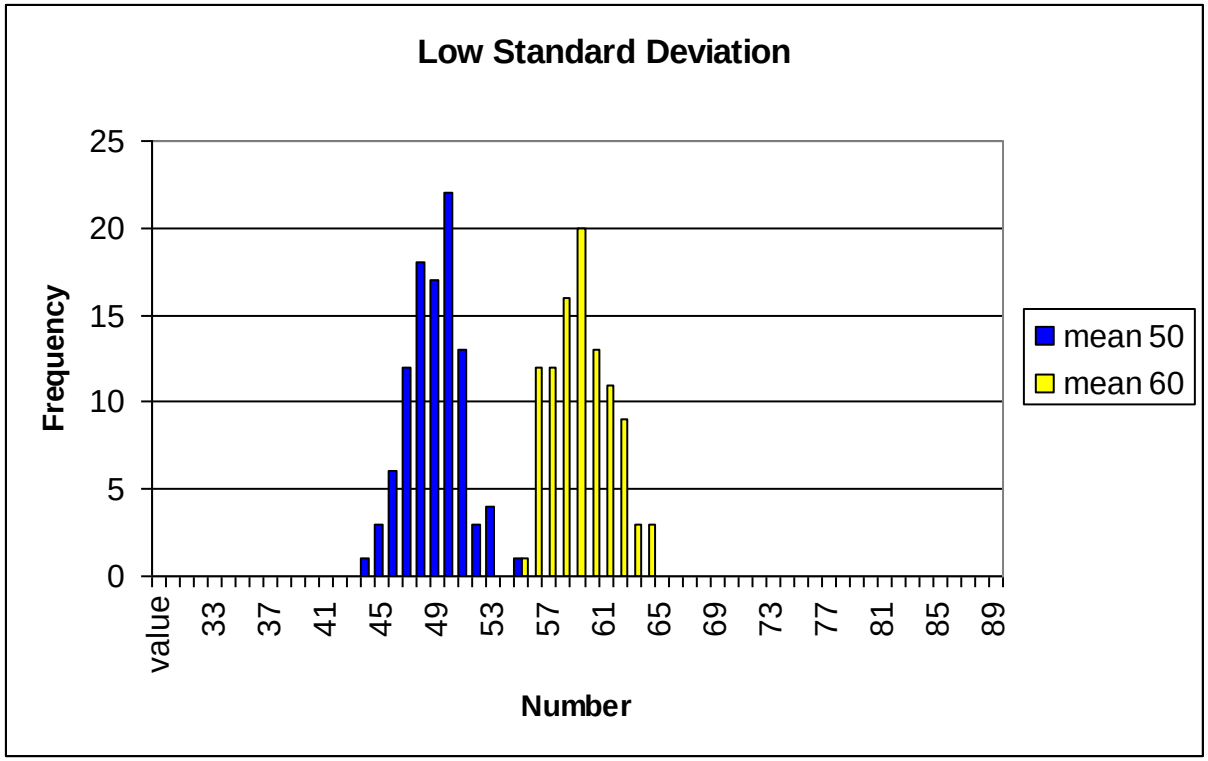
The size of the sample:

- What can we conclude if we have one treated village and one non treated village?
- What can we conclude if we give malaria medicine (IPT) to one classroom and not the other?
- Even though we have a large class size?
- What matters is the effective sample size i.e. the number of treated units and control units (e.g. class rooms). What is the unit in the case of IPT given in the classroom?

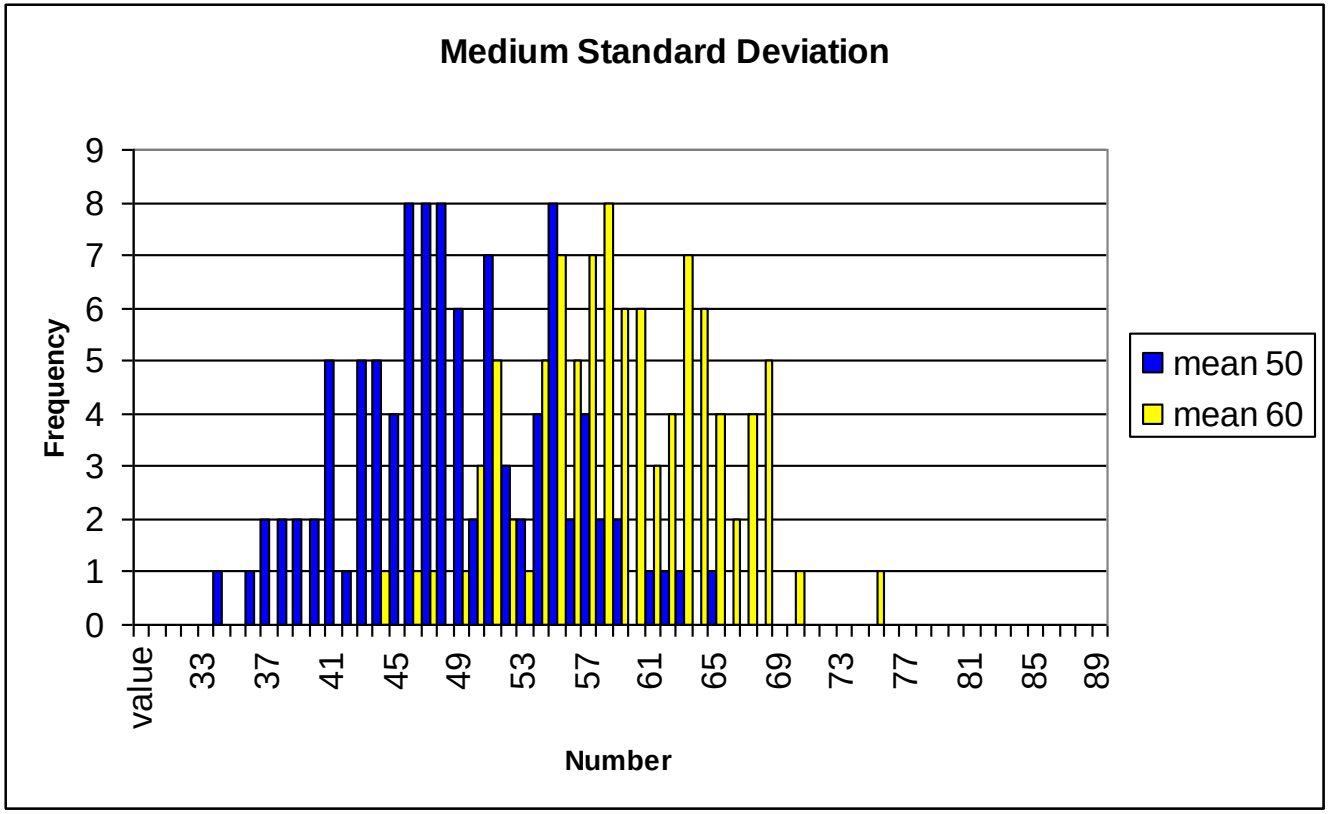
The variability in the outcome we try to measure:

- If there are many other non-measured things that explain our outcomes, it will be harder to say whether the treatment really changed it.

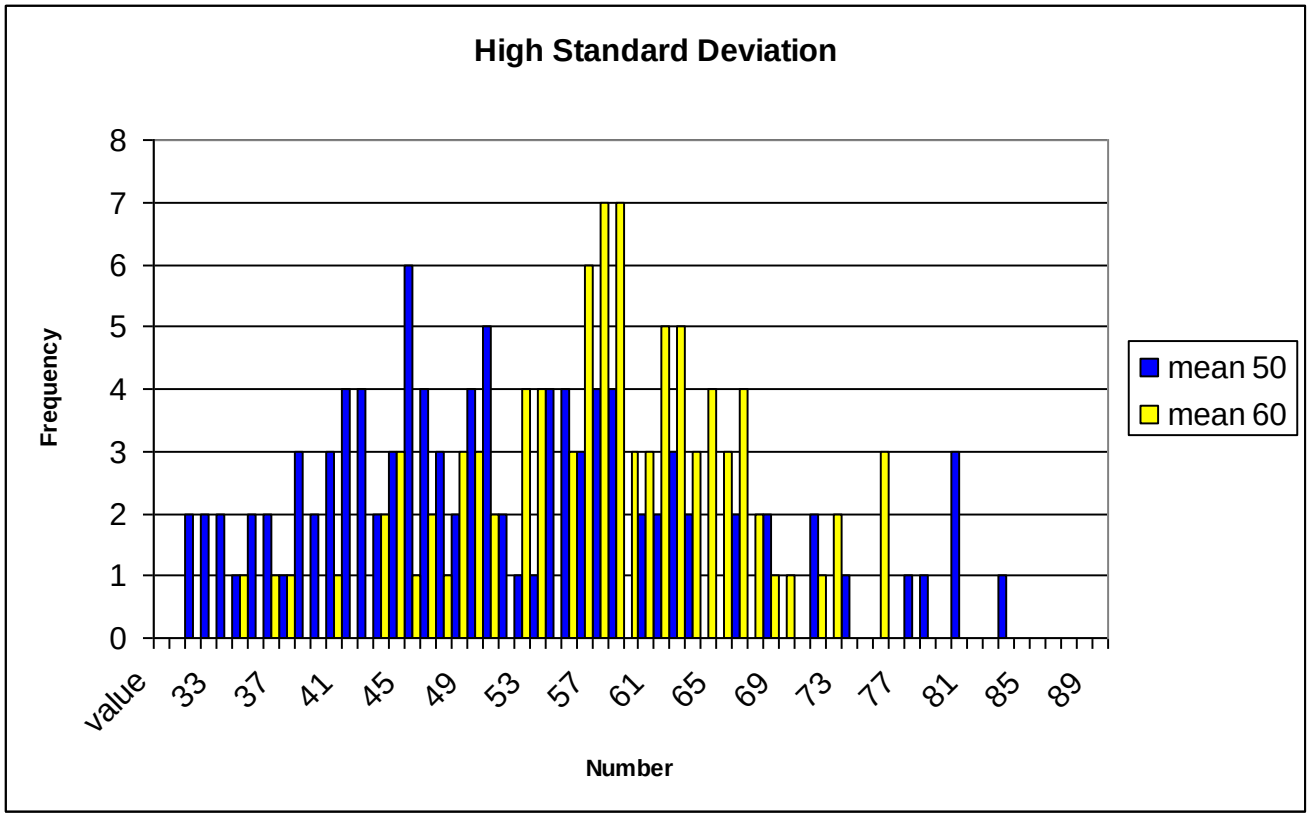
When the Outcomes are Very Precise



Less Precision



What can we Conclude?



Confidence Intervals

- The estimated effect size (the difference in the sample averages) is valid only for our sample. Each sample will give a slightly different answer. How do we use our sample to make statements about the overall population?
- A 95% confidence interval for an effect size tells us that, for 95% of all samples that we could have drawn from the same population, the estimated effect size would fall within this interval.
- The Standard error (se) of the sample estimate captures both the size of the sample and the variability of the outcome (it is larger with a small sample and with a variable outcome)
- Rule of thumb: a 95% confidence interval is roughly the effect plus or minus two standard errors.

Hypothesis Testing

Often we are interested in testing the hypothesis that the effect size is equal to zero (we want to be able to reject the hypothesis that the program had no effect)

We want to test:

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

Against:

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

Two Types of Mistakes

- First type of error : Conclude that there is an effect, when in fact there is 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 purpose, you want to be very confident in the answer you give: the level will be set fairly low.

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

Relation with Confidence Intervals

- If zero does not fall inside the 95% confidence interval of the effect size we measured, then we can be 95% sure that the effect size is not zero

- 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

Second type of error: you fail to reject that the program had no effect, when in 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 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 size
- One minus the power is the probability to be disappointed….

Calculating Power

- When planning an evaluation, with some preliminary investigation we can calculate the minimum sample we need to get to:
  - Test a pre-specified 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 at 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 is 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 outcome 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
- 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
Bed net distribution	Health clinics
IPT	Schools
Social support	Family

Reason 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 (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 members of a village interact with each other
- The sample size needs to be adjusted for this correlation
- The more correlation between outcomes within the cluster, the more we need to adjust the standard errors

A Simple Estimation Framework

$$Y_{ij} = \gamma_0 + \gamma_1 W_j + \mu_j + e_{ij}$$

$$e_{ij} \sim N(0, \sigma^2), \mu_j \sim N(0, \tau^2)$$

- *i = 1,…,n persons per cluster and j = 1,…J clusters*
- *W_j is an indicator variable that represents treatment*
- *μ_j is the effect associated with each cluster*
- *e_{ij} is the error associated with each individual*

A Simple Estimation Framework (cont.)

The estimated effect size:

$$\hat{\gamma} = \overline{Y_T} - \overline{Y_C}$$

We can derive the variance of the estimator:

$$Var(\hat{\gamma}) = \frac{4(\tau^2 + \frac{\sigma^2}{n})}{J}$$

- *Where n is number of individuals per cluster and J is the number of clusters*

A Simple Estimation Framework (cont.)

We also talk about the intra-cluster or intra-class correlation:

$$\rho = \frac{\tau^2}{\tau^{2+}\sigma^2}$$

Which enables us to rewrite the variance of the estimator:

$$Var(\hat{\gamma}) = \frac{4(\rho + \frac{1-\rho}{n})}{J}$$

Example of Group Effect Multipliers

Intra-Class Correlation	Randomized Group Size 10	50	100	200
0.00	1.00	1.00	1.00	1.00
0.02	1.09	1.41	1.73	2.23
0.05	1.20	1.86	2.44	3.31
0.10	1.38	2.43	3.30	4.57

Implications

- It is extremely important to randomize an adequate number of groups
- Often the number of individuals within groups matter less than the number of groups
- Think that 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!!!!

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: implications for cost
  - Can be used to stratify and form subgroups
- To compute power with a baseline:
  - You need to know the correlation between subsequent measures 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

Control Variables

If we have additional relevant variables (e.g. village population, block where the village is located, etc.) we can also control for them

What matters now for power is the residual variation 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: usually 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 ensures 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 anti-mosquito spray program
  - Agroclimatic and associated epidemiologic factors are controlled for
  - The “common district government effect” disappears.

The Design Factors that Influence Power

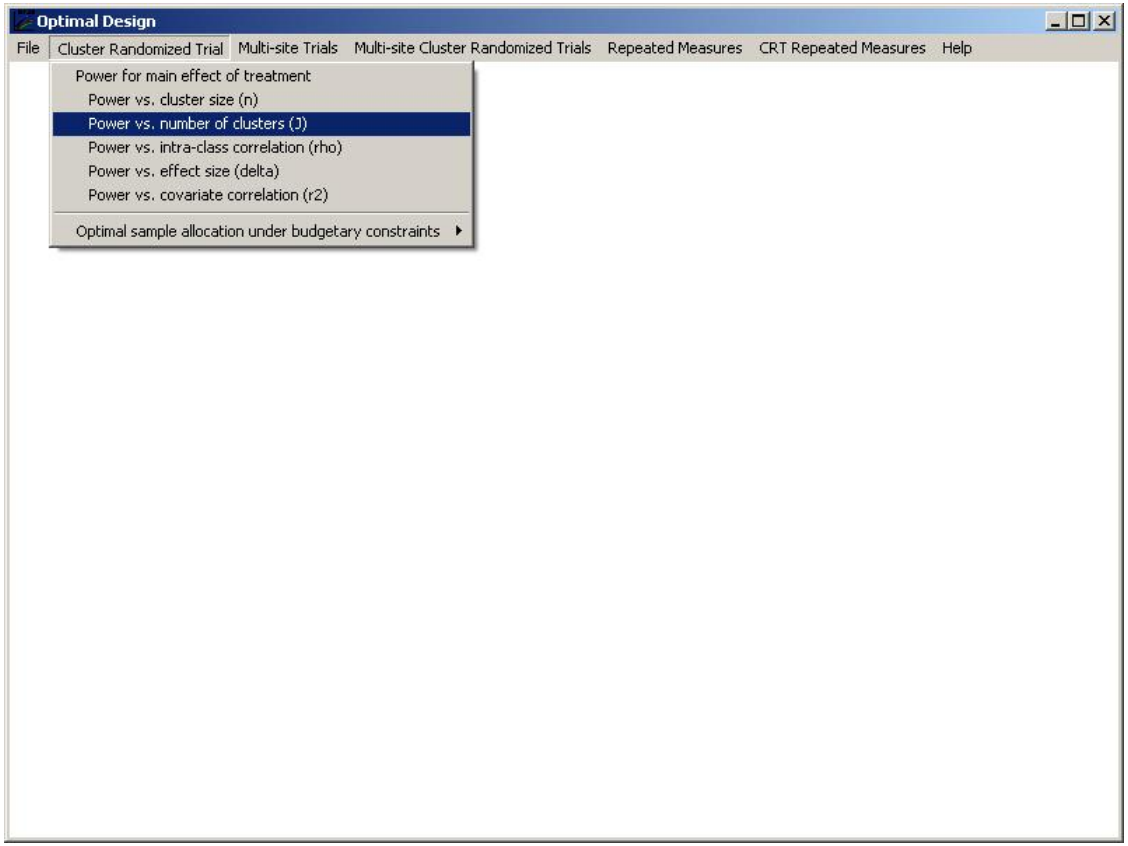
- Clustered design
- Availability of a baseline
- Availability of control variables, and stratification.
- The type of hypothesis that is being tested.

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? (e.g. encouragement design?)

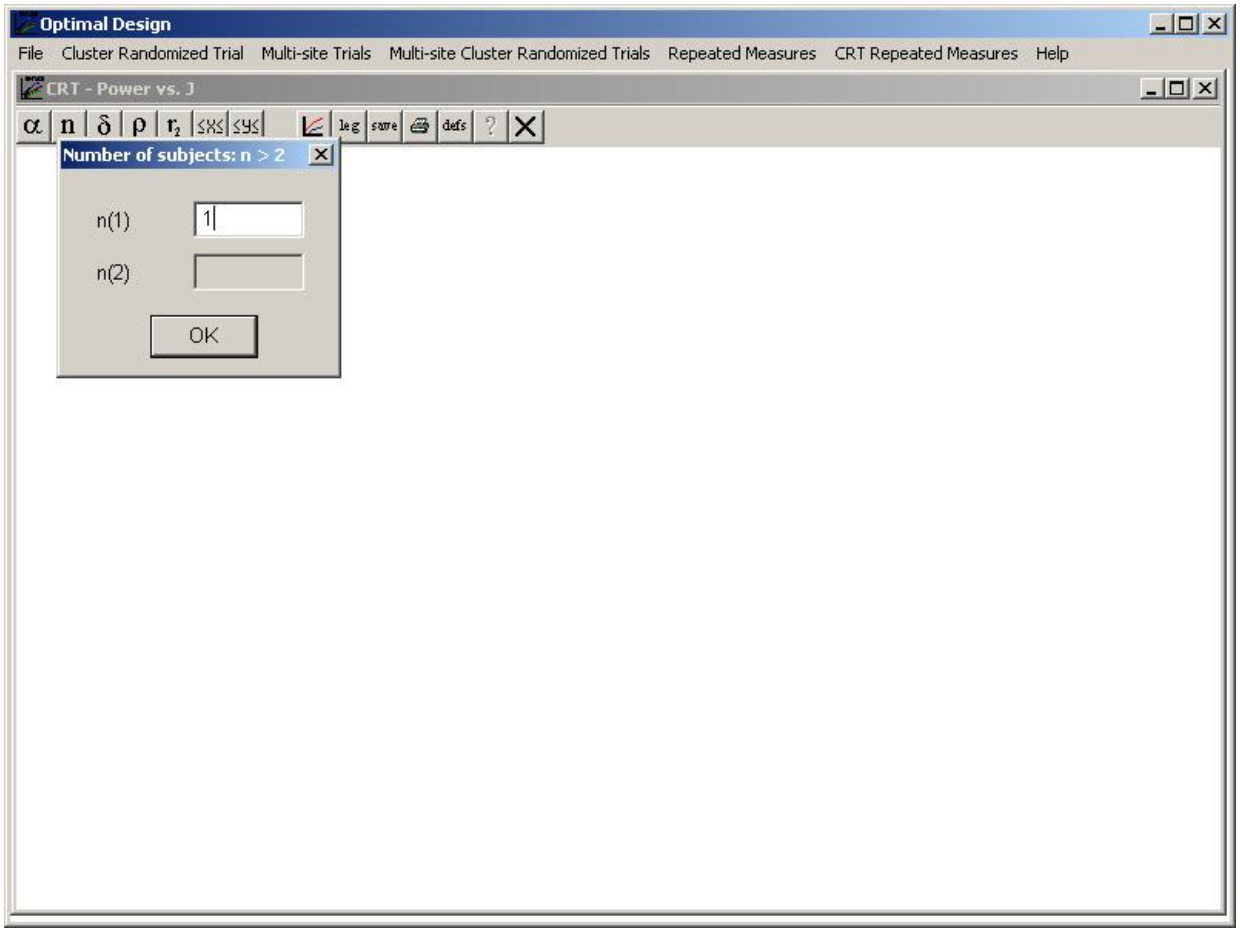
Power Calculations Using the OD Software

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



Cluster Size

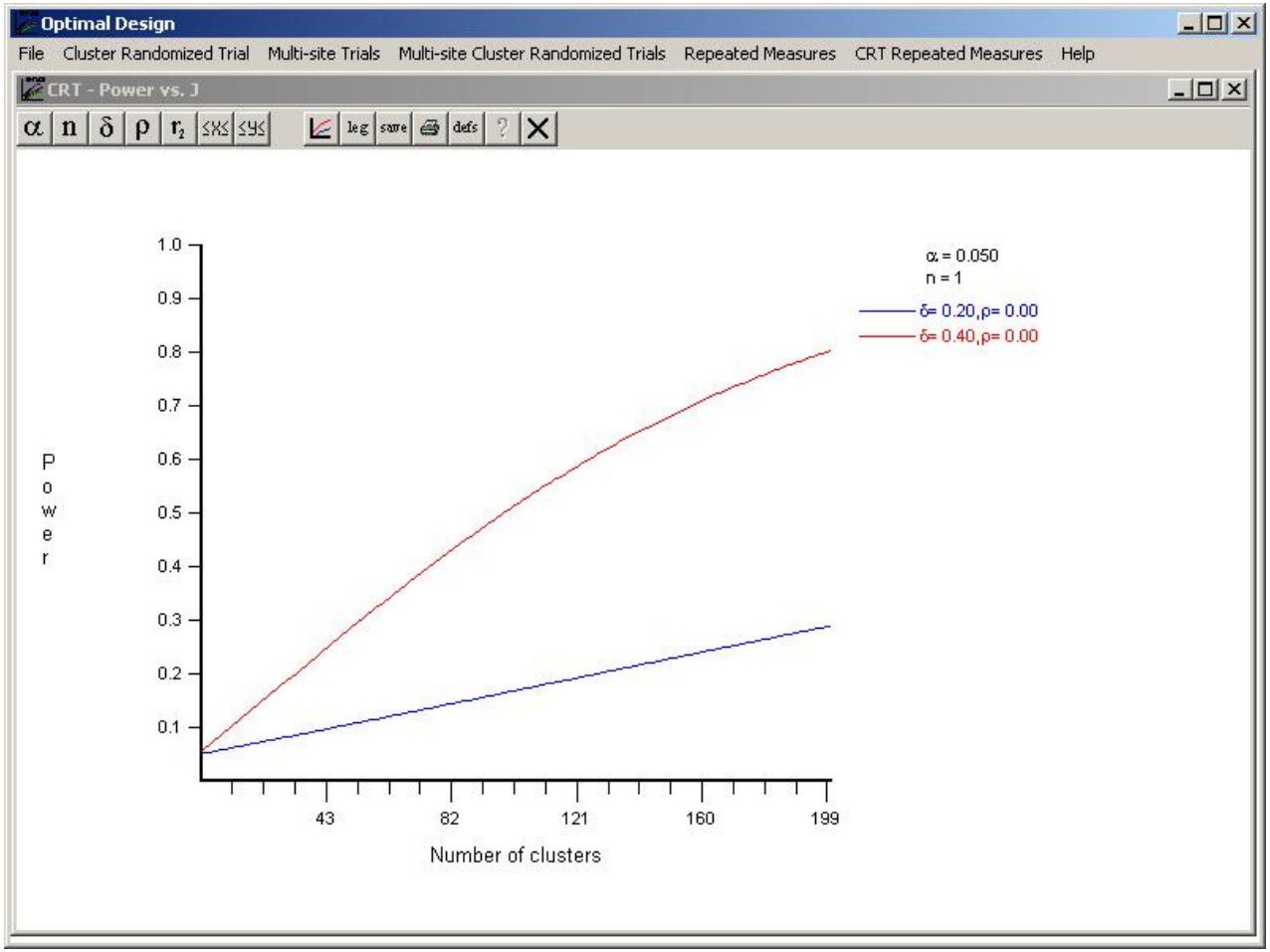
- Choose cluster size



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



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)