Abdul Latif Jameel Poverty Action Lab Executive Training: Evaluating Social Programs
Spring 2009

Planning sample size for randomized evaluations

Abdul Latif Jameel
Poverty Action Lab

Planning an evaluation

- Today's Question:

  How large does the sample need to be to credibly detect a given treatment effect?
- What does credibly mean?

  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?

Sample size

- Important determinants of sample size
  - How big an effect size are we looking for?
  - How noisy is the outcome measure?
  - Do we have a baseline?
  - Are individual responses correlated with each other?
  - Design of the experiment: stratification, control variables, baseline data, group v. individual level randomization

Outline

I. Hypothesis testing

II. Type I and Type II Errors

III. Standard errors and significance

IV. Power

V. Effect size

VI. Factors that influence power

Hypothesis testing: Simple intuition I

- Professional gambler, claims she can get heads most of the time with a fair coin
  - One toss: “H”
    - Any inference?
  - Five tosses: H,H,T,H,H
    - Any inference?
  - Twenty Tosses:
    - T,H,T,H,T,H,H,H,T,H,T,H,T,H,H,T,H,T,H,H
    - (12 Head, 8 Tails)
  - One hundred tosses
    - 61 Heads, 39 Tails
  - One thousand tosses
    - 609 Heads, 391 Tails

Very simple intuition II

- Second gambler, 1,000 tosses,
  - Observe 530 Heads, and 470 tails.
- Can we reject claim that he obtains H 70% of the time? (e.g., 20% more than 50%)?
- Can we reject claim that he obtains H 54% of the time (e.g., 4% more than 50%)?

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

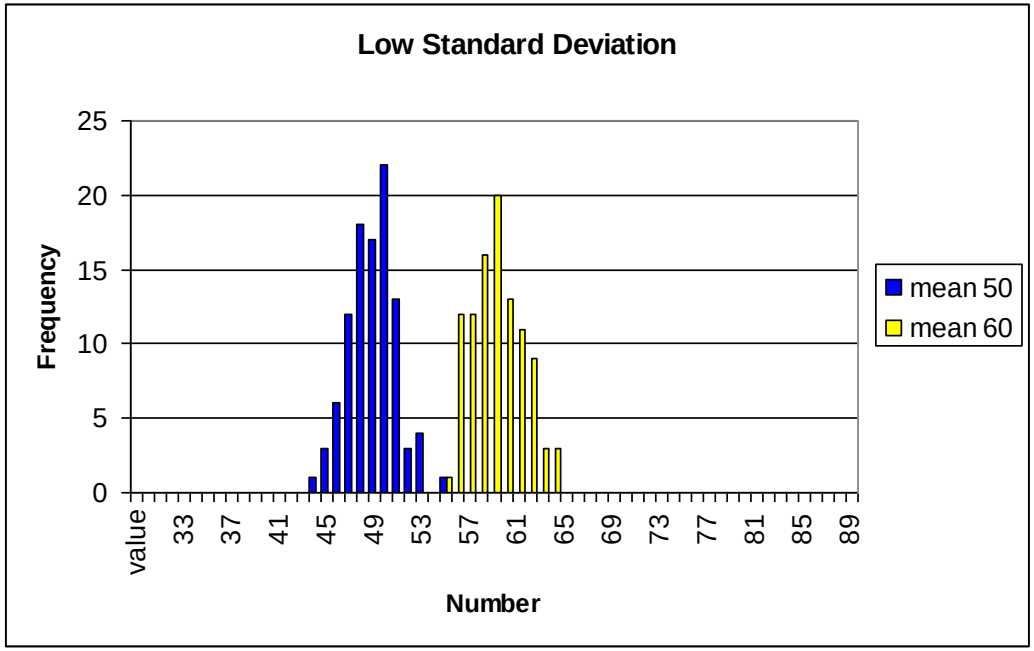
  = Effect size

  - For example: mean of the number of wells in villages with female leaders vs mean of the number of wells in villages with male leaders

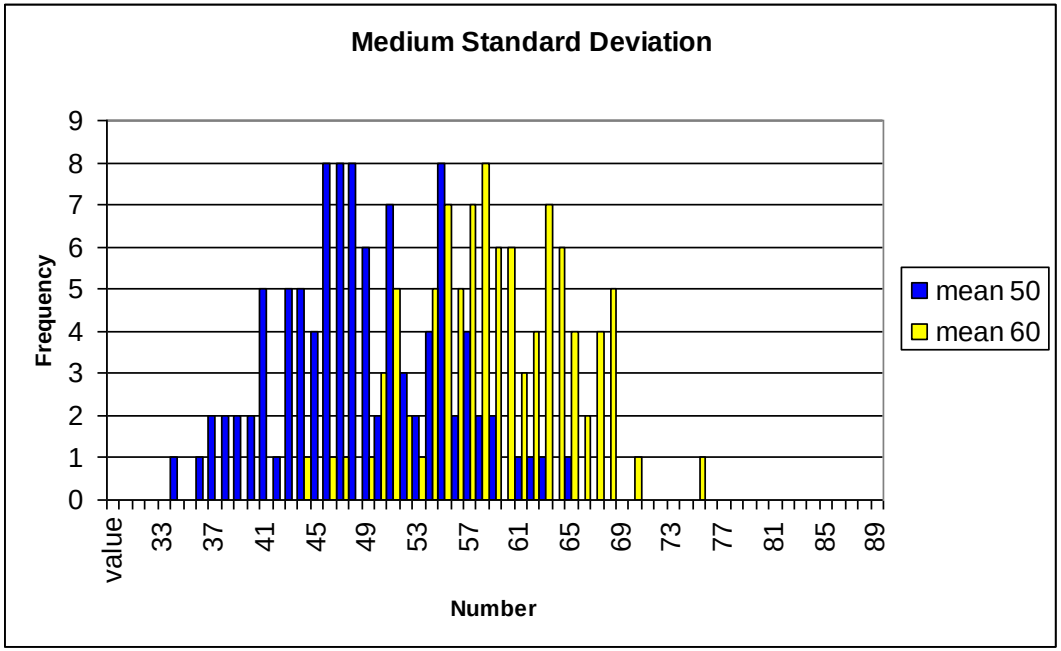
Estimation

- But we do not observe the entire population, just a sample.
  - In each village of the sample, there is a given number of wells. It is more or less close to the mean in the population.
- 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

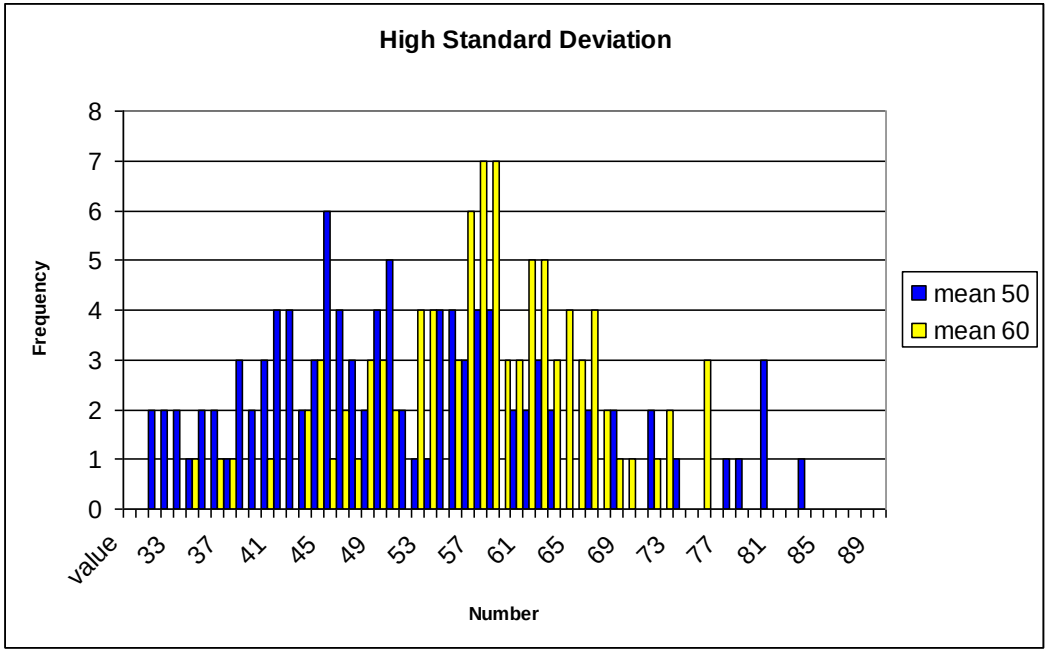
A tight conclusion



Less precision



Can we conclude anything?



Very simple intuition

- Sample Size Matters:
  - The more tosses we have, the better able we are to understand the true probability of heads
- The hypothesis matters
  - The more fine (or more precise) the effect size we want to detect, the more tosses we need
- Variability of the outcome matters
  - The more “noisy” it is, the harder it is to measure effects

Intuition: Confidence intervals

- We measure the length of 100 randomly selected infants, and find an average length of 53 cm?
- How precise is that estimate? Could it be, if we measure all the infants, we would in fact find an average of 54 cm? Or 60 cm?
- Confidence interval: given some data, a sense of how precise our estimate is
- A confidence interval of 50-56 says that with 95% probability, the true average length lies between 50 and 56.
- Approximate interpretation: “We know the point estimate of 53 isn’t exactly correct, but its close…how close? Well, it’s very likely that the true answer is between 50 and 56.

Confidence intervals

- The goal is to figure out the true effect of the program
- From our sample, we get an *estimate* of the program effect
- What can we learn about the *true program effect* from the *estimate?*
- A 95% confidence interval for an estimate tells us that, with 95% probability, the true program effect lies within the confidence interval
- The Standard error (se) of the estimate in the sample captures both the size of the sample and the variability of the outcome (it is larger with either a small sample or with a volatile outcome)
- Rule of thumb: a 95% confidence interval is roughly the effect plus or minus two standard errors.

Confidence intervals

- Example 1:
  - Sampled women Pradhans have 7.13 years of education
  - Sampled male Pradhans have 9.92 years of education
  - The difference is 2.59 with a standard error of 0.54
  - The 95% confidence interval is [1.53; 3.64]
- Example 2:
  - Control children have an average test score of 2.45
  - Treated children have an average test score of 2.50
  - The difference is 0.05, with a standard error of 0.26
  - The 95% confidence interval is [-0.55;0.46]

Hypothesis testing

- Often we are interested in testing the hypothesis that the effect size is equal to zero:
- We want to test the *null hypothesis (H_0)*:

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

Against the *alternative hypothesis (H_a)*:

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

(other possible alternatives: $H_a>0$, $H_a<0$, $H_a>2$).

- Hypothesis testing asks: when can I reject the null in favor of the alternative?

Outline

I. Hypothesis testing

II. **Type I and Type II Errors**

III. Standard errors and significance

IV. Power

V. Effect size

VI. Factors that influence power

Two types of mistakes

- Type I error: Conclude that there is an effect, when in fact there are no effect.

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

Example: Female Pradhan's year of education is 7.13, and Male's is 9.92 in our sample. Do female Pradhan have different level of education, or the same?

If I say they are different, how confident am I in the answer?

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

Common level of significance: 0.05, 0.01, 0.1.

Two types of mistakes

- Type II error: you fail to reject that 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 (higher power is better since I am more likely to have an effect to report, if there is one.)
  - Power is a planning tool. It tells me how likely it is that I find a significant effect for a given sample size, if there is one.

Example: If I run 100 experiments, in how many of them will I be able to reject the hypothesis that women and men have the same education at the 5% level, if in fact they are different?

Intuition

		YOU CONCLUDE	
		Effective	*No Effect*
THE TRUTH	*Effective*	☺	**Type II Error** (power) ☹
	No Effect	**Type I Error** (size) ☹	☺

Outline

I. Hypothesis testing

II. Type I and Type II Errors

III. Standard errors and significance

IV. Power

V. Effect size

VI. Factors that influence power

Testing equality of means

We have $\hat{\beta}$

(1) our estimate of the program effectiveness.

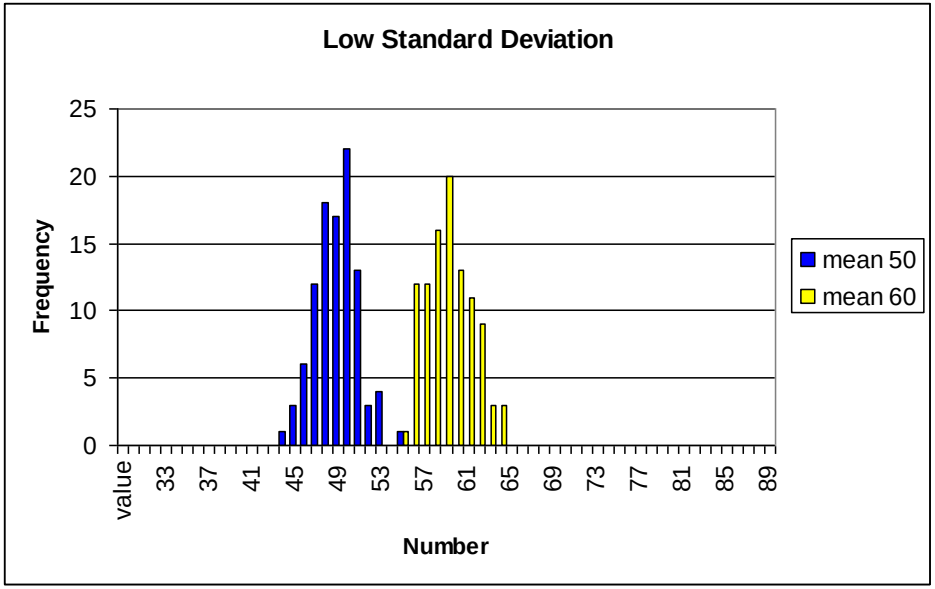
For example

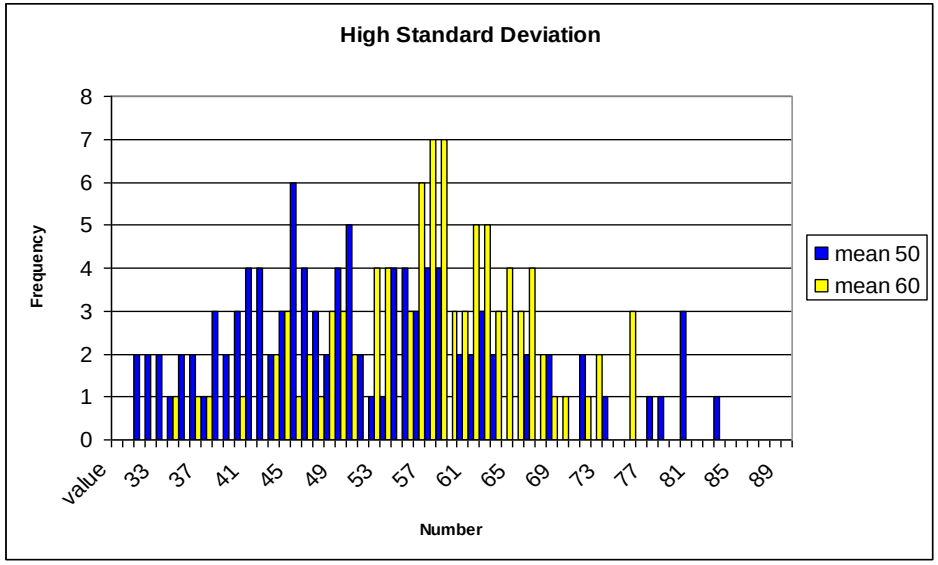
= Average Treated Test Score– Average Control Test Score

(2) An estimate of the "standard error" of $\hat{\beta}$, which measures how precise our estimate is. (The same thing used to compute confidence intervals).

(Depends on the variability of $\hat{\beta}$ and sample size)

Standard error intuition





Testing equality of means

We use $t = \dfrac{\hat{\beta}}{se(\beta)}$

- So if $t > 1.96$, we reject the hypothesis of equality at a 5% level of confidence (5% chance there is in fact no difference)
- It $t < 1.96$, we *fail to reject* the hypothesis of equality at a 5% level of confidence
- Example of Pradhan's education:
  - Difference: 2.59
  - Standard error: 0.54
  - We definitely reject equality at 5% level.

Outline

I. Hypothesis testing

II. Type I and Type II Errors

III. Standard errors and significance

IV. Power

V. Effect size

VI. Factors that influence power

Calculating power

- When planning an evaluation, with some preliminary research we can calculate the minimum sample we need to get to:
  - Test a pre-specified null hypothesis (e.g. treatment effect 0)
  - For a pre-specified significance level (e.g. 0.05)
  - Given a pre-specified effect size (e.g. 0.2 standard deviation of the outcomes of interest).
  - 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 H_o is in fact false (e.g. the treatment effect is not zero), we will be able to reject it.
- 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 conventionally set at 5% The lower it is, the larger the sample size needed for a given power
The mean and the variance of the outcome in the comparison group	From a small survey in the same or a similar population 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? Rationale: If the effect is any smaller than this, then it is not interesting to distinguish it from zero

Outline

I. Hypothesis testing

II. Type I and Type II Errors

III. Standard errors and significance

IV. Power

V. Effect size

VI. Factors that influence power

Picking an effect size

- What is the smallest effect that should justify the program being adopted
  - Cost of this program vs the benefits it brings
  - Cost of this program vs 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
- ***NOT*** : “expected” effect size

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 Standardized effect size is the effect size divided by the standard deviation of the outcome

  δ = effect size/St.dev.
- Common effect sizes:

$\delta = 0.20$ (small) $\delta = 0.40$ (medium) $\delta = 0.50$ (large)

Standardized effect sizes

An effect size of ..	Is considered . . .	.It means that . . .
0.2	small-modest	The average member of the intervention group had a better outcome than **58 percent** of the members of the control group.
0.5	modest-large	The average member of the intervention group had a better outcome than **69 percent** of the members of the control group.
0.8	large	The average member of the intervention group had a better outcome than **79 percent** of the members of the control group.

Outline

I. Hypothesis testing

II. Type I and Type II Errors

III. Standard errors and significance

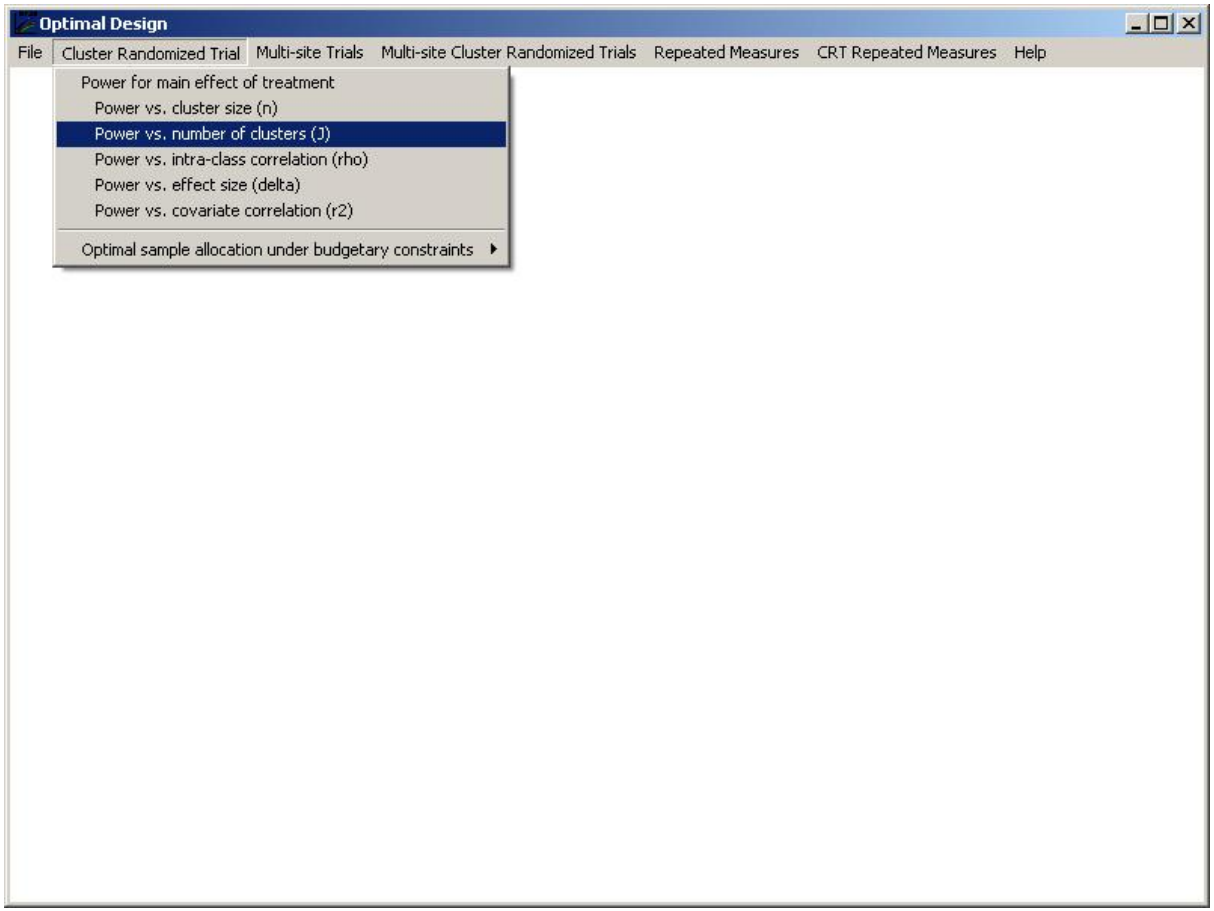
IV. Power

V. Effect size

VI. Factors that influence power

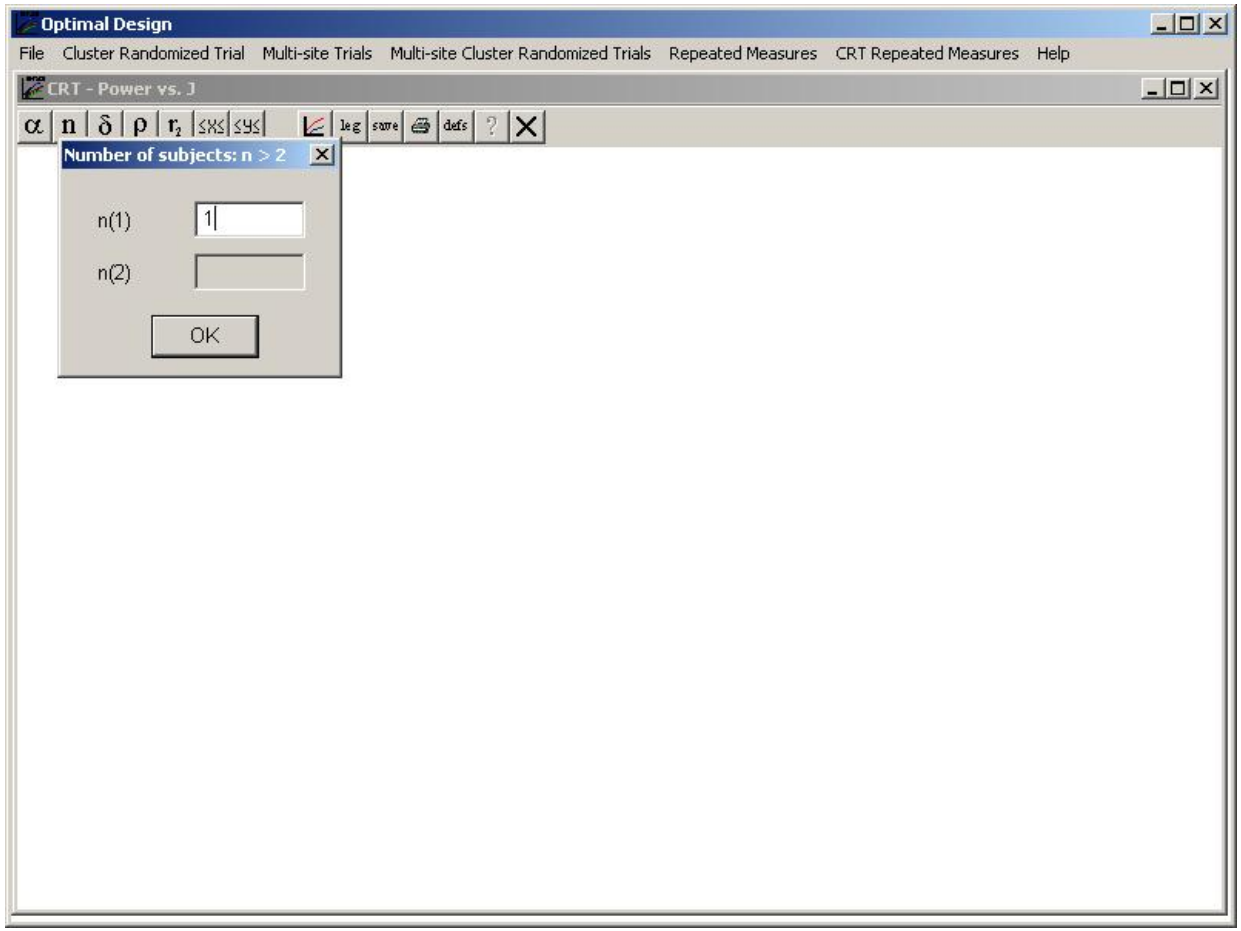
Power calculations using the OD software

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



Cluster size

- Choose cluster with 1 unit (non-clustered design)

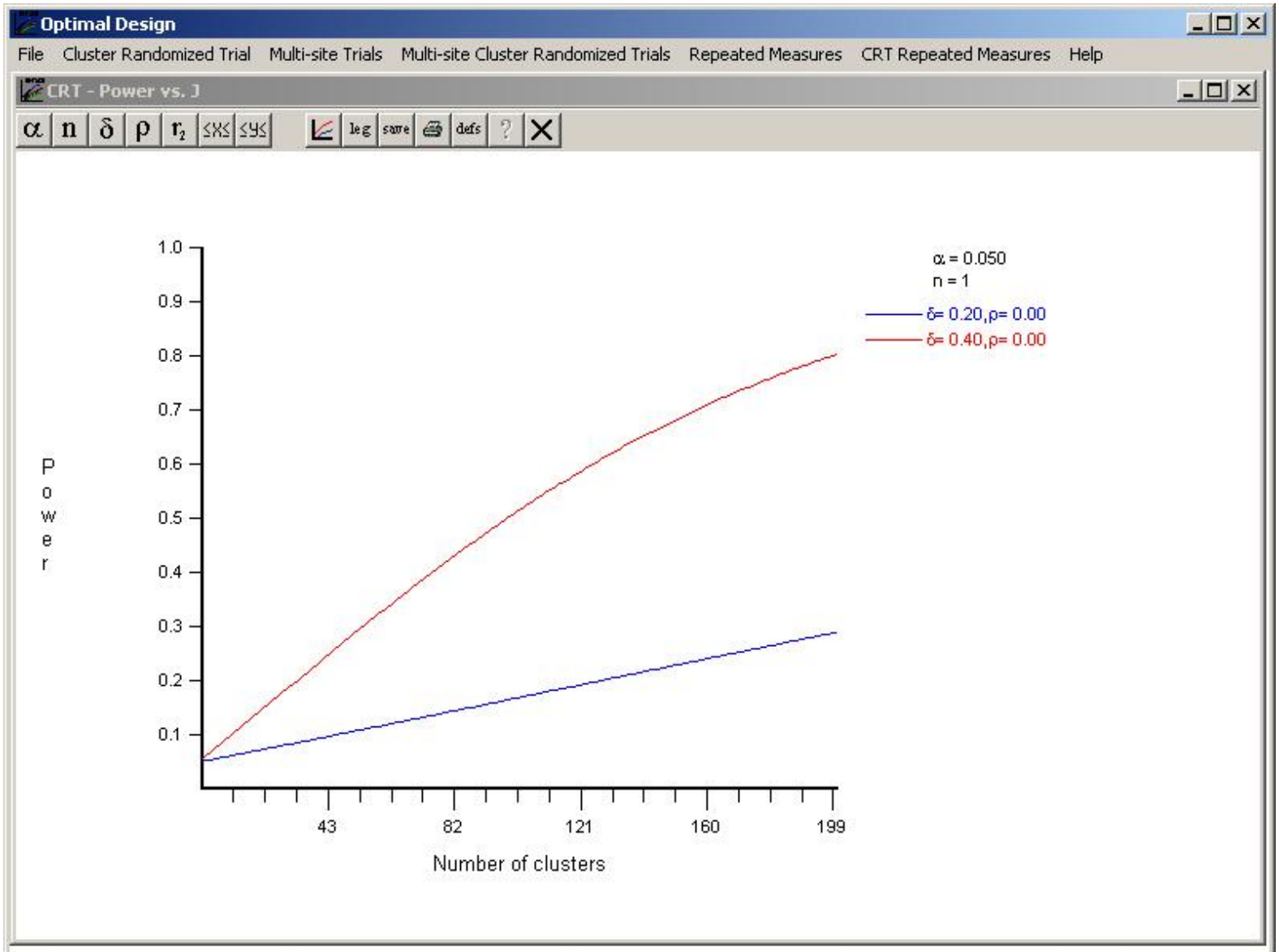


Choose significance level and treatment effect

- Pick α
  - Normally you pick 0.05
- Pick δ
  - Can experiment with 0.20 (small effect size)

- You obtain the resulting graph showing power as a function of sample size.

Power and sample size



The design factors that influence power

1. The level of randomizaion - clustered design
2. Availability of a baseline
3. Availability of control variables, and stratification.
4. The type of hypothesis that is being tested.

Intuition: Clustered design

- You want to know how close the upcoming national elections will be
- Method 1: Randomly select 50 people from entire Indian population
- Method 2: Randomly select 5 families, and ask ten members of each family their opinion

Intuition: Clustered design

- If the response is correlated within a group, you learn less information from measuring multiple people in the group
- It is more informative to measure *unrelated* people
- Measuring *similar* people yields less information

Clustered design

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

Examples:

PROGRESA	Village
Gender Reservations	Panchayats
Flipcharts, Deworming	School
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 affects an entire classroom (e.g. textbooks, 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 Panchayats share a common history
  - All students share a schoolmaster
  - The program affect all students at the same time.
  - The member of a village interact with each other
- We call ρ (rho) the correlation between the units within the same cluster

Values of ρ (rho)

- Like percentages, ρ must be between 0 and 1
- When working with clustered designs, a lower ρ is more desirable
- It is sometimes low, 0, .05, .08, but can be high:

Madagascar Math+language	0.5
Busia, Kenya Math+language	0.22
Udaipur, India Math+language	0.23
Mumbai, India Math+language	0.29
Vadodara, India Math+language	0.28
Busia, Kenya Math	0.62

Implications for design and analysis

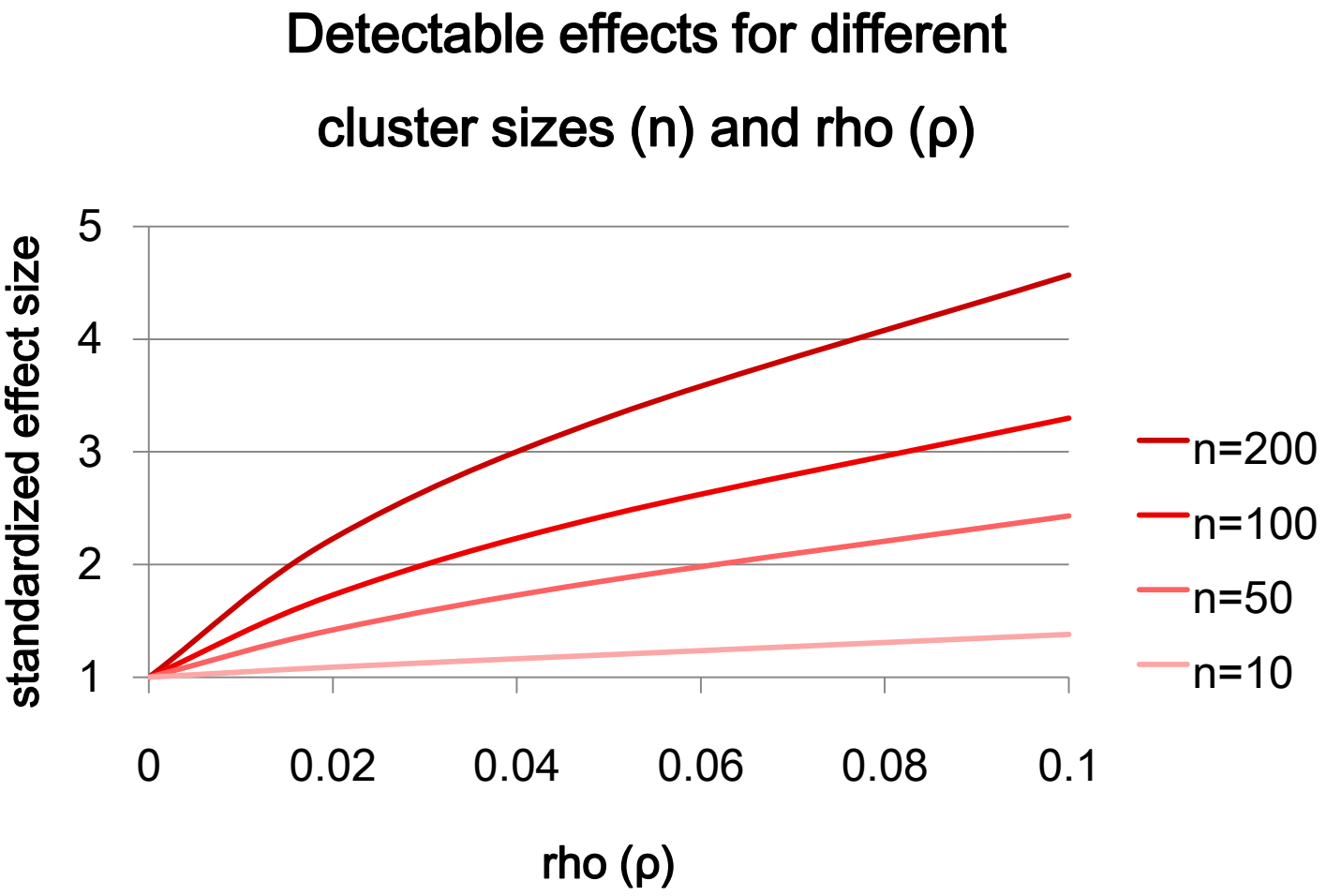
- Analysis: The standard errors will need to be adjusted to take into account the fact that the observations within a cluster are correlated.
- Adjustment factor (design effect) for given total sample size, clusters of size m, intra-cluster correlation of r, the size of smallest effect we can detect increases by $\sqrt{1+\rho*(m-1)}$ compared to a non-clustered design
- Design: We need to take clustering into account when planning sample size

Example: detectable treatment size vs. rho (ρ)

Intraclass	Randomized Group Size			
Correlation (ρ)	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

i.e. When clusters have 100 people, detectable treatment size more than triples ...

Example: detectable treatment size vs. rho (ρ)



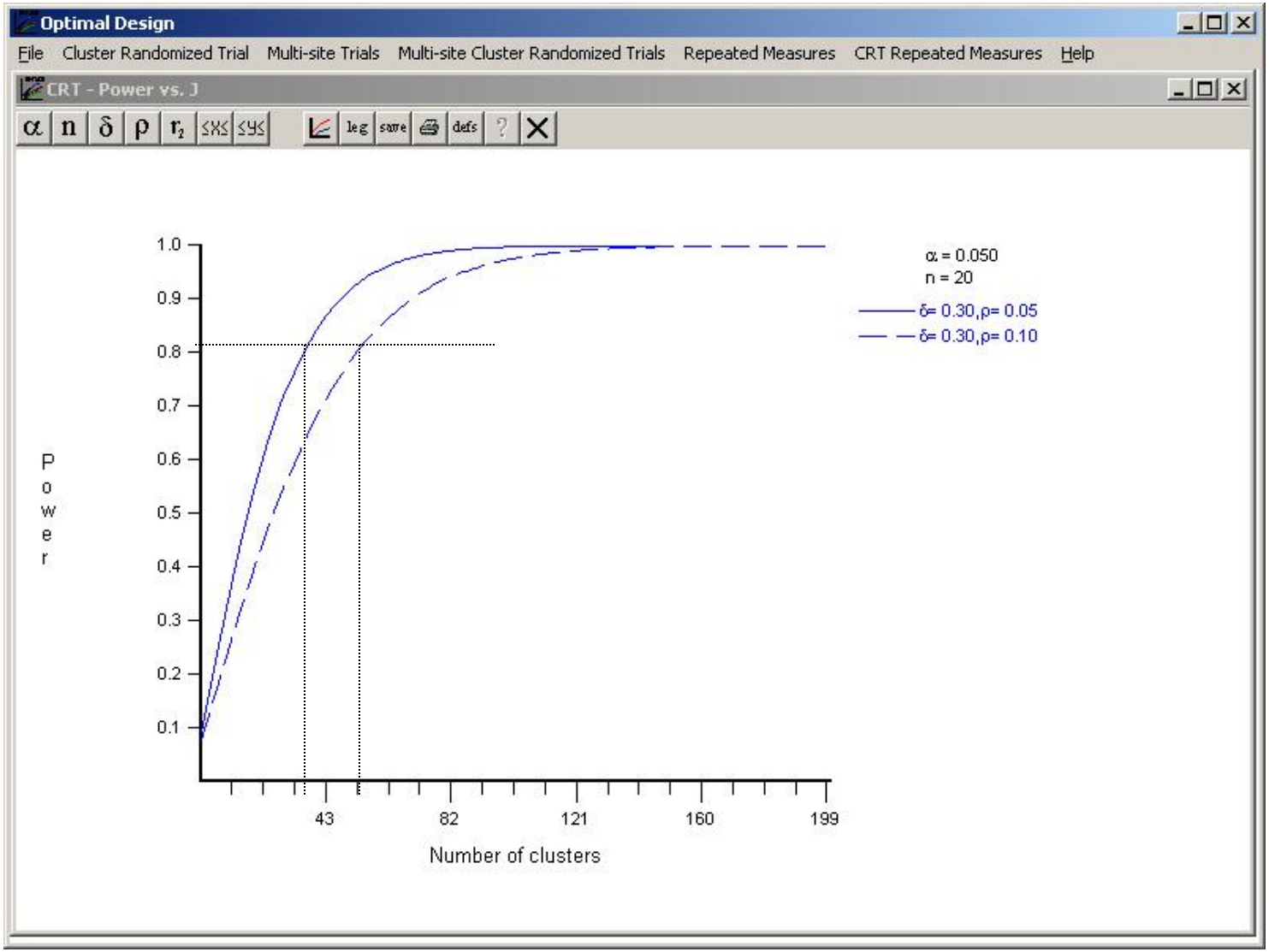
Implications

- If experimental design is clustered, we now need to consider ρ when choosing a sample size (as well as the other effects)
- It is extremely important to randomize an adequate number of groups
- Often the number of individuals within groups matter less than the **total number of groups**

Choosing the number of clusters with a known number of units

- Example: Randomization of a treatment at the classroom level with 20 students per class:
  - Choose other options as before
  - Set the number of students per school (e.g. 20)
  - set ρ

Power Against number of clusters with 20 students per cluster



38 vs. 53 clusters needed for 80% power

Choosing the number of clusters when we can choose the number of units

- To chose how many Panchayats to survey and how many villages per Panchayats to detect whether water improvement are significantly different for women and men
- Mean drinking water facilities built or repaired in unreserved GPs: 14.7
- Standard deviation: 19
- ρ: 0.07

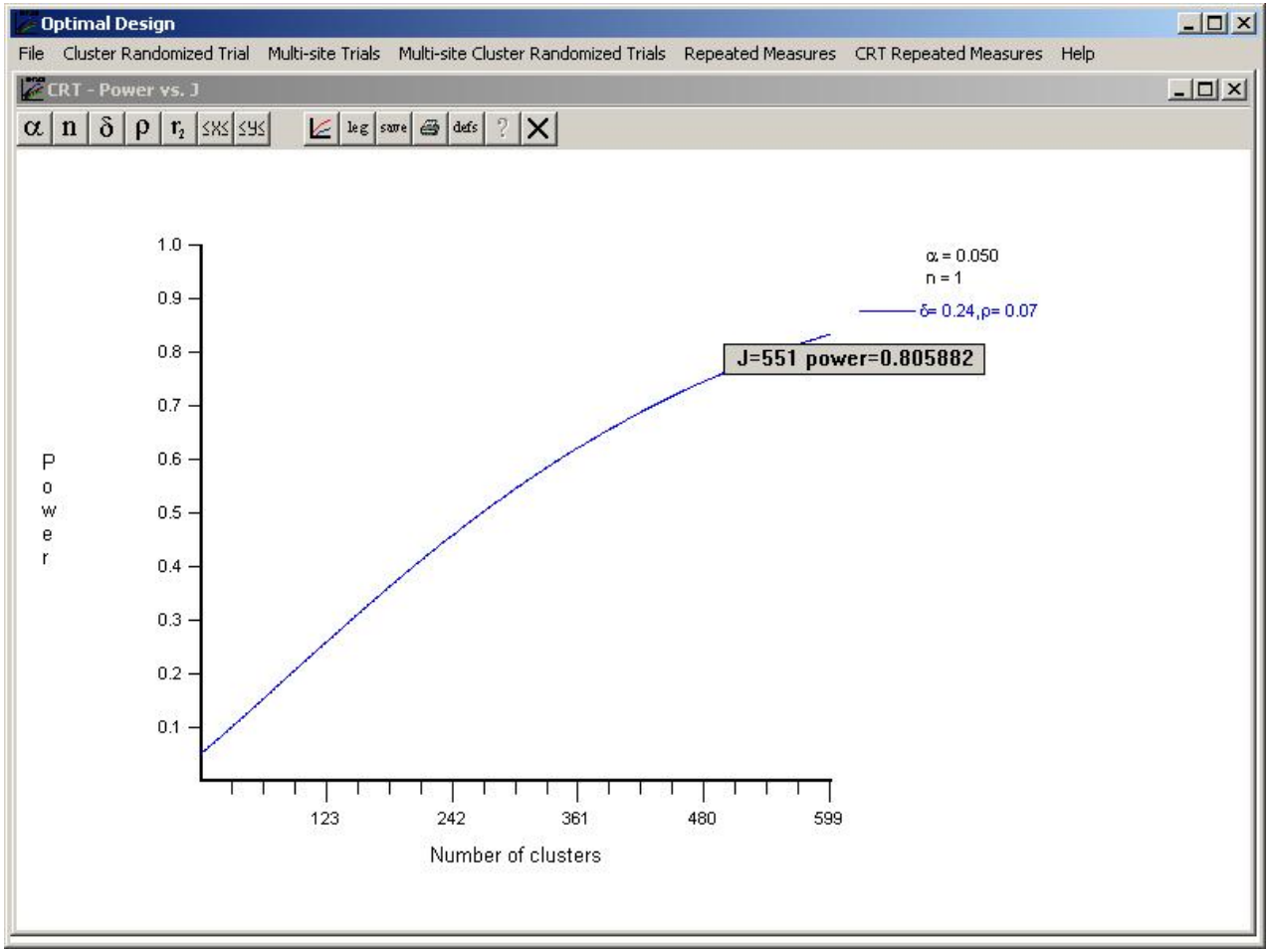
Calculating effect size

- Mean drinking water facilities built or repaired in unreserved GPs: 14.7
- Standard deviation: 19
- We want to detect at least a 30% increase
- 30% of 14.7 is 4.41
- 4.41/19=.23 standard deviations
- delta = 0.23
- We look for a power of 80%

Minimum number of GP's, fix villages per GP

- We search for the minimum number of GP we need if we survey 1 village per GP:
  - Answer: 553

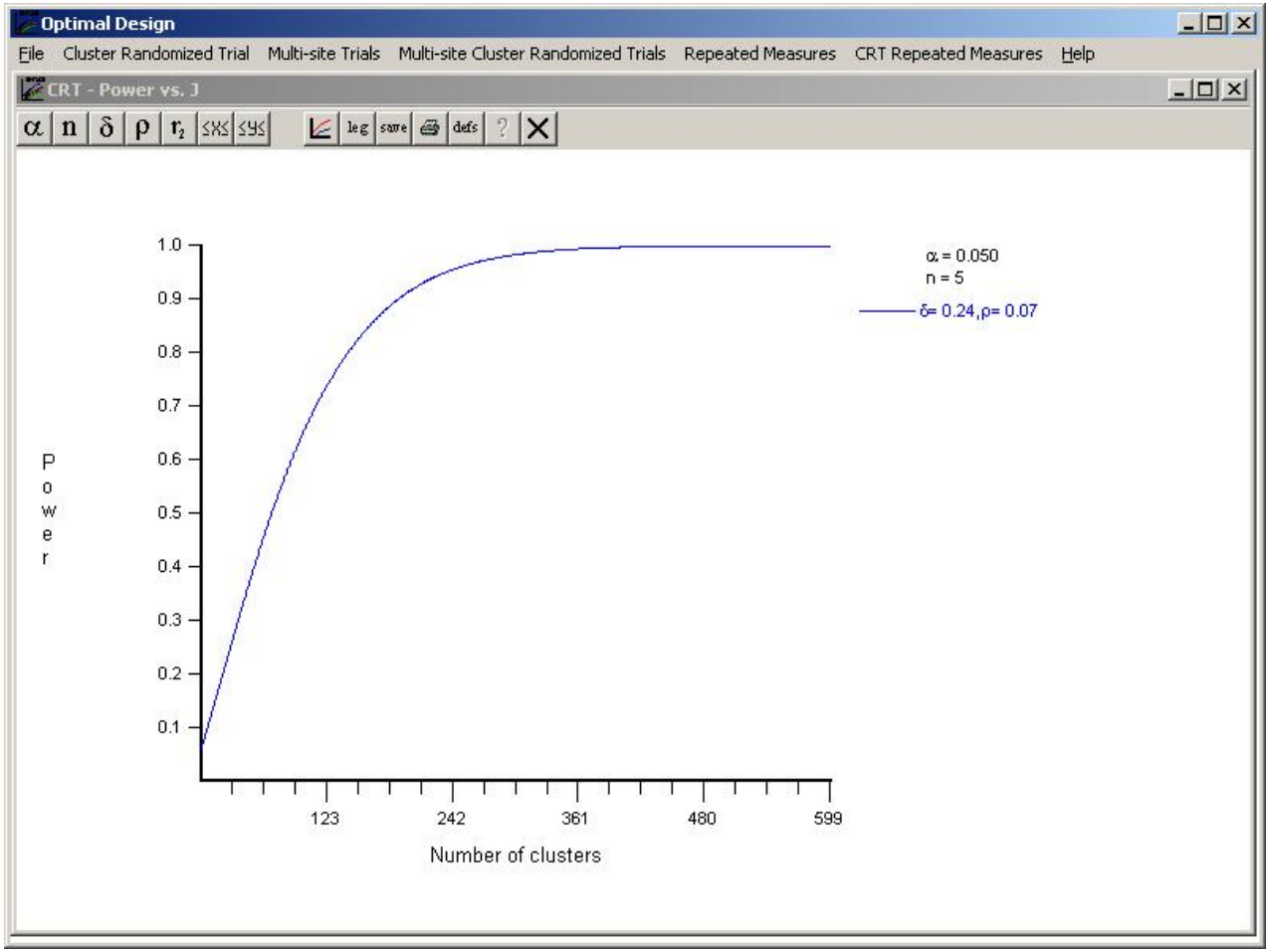
Number of clusters for 80% power



Minimum number of GP's, fix villages per GP

- We search for the minimum number of GP we need if we survey 1 village per GP:
  - Answer: 553
- We search for the minimum number of GP if we survey 2, 3, 4, etc… village per GP

Power against number of clusters with 5 villages per panchayat



Minimum number of GP's, fix villages per GP

- We search for the minimum number of GP we need if we survey 1 village per GP:
  - Answer: 553
- We search for the minimum number of GP if we survey 2, 3, 4, etc… village per GP
- For each combination, we calculate the number of villages we will need to survey, and the budget.

What sample do we need?

Exercise A			
Power: 80%			
# of village per GP	# of GP's	total # of villages	Total Cost (man days)
1	553	553	3041.5
2	297	594	2673.0
3	209	627	2612.5
4	162	648	2592.0
5	141	705	2749.5
6	121	726	2783.0
7	107	749	2835.5
8	101	808	3030.0

The design factors that influence power

1. Clustered design
2. **Availability of a Baseline**
3. Availability of Control Variables, and Stratification.
4. The type of hypothesis that is being tested.

Availability of a baseline

- A baseline has two main uses:
  - Allows you to check whether control and treatment group were the same or different before the treatment
  - **Reduces 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**
- To compute power with a baseline:
  - You need to know the correlation between two subsequent measurements of the outcome (for example: correlation between pre and post test score in school).
  - The stronger the correlation, the bigger the gain.
  - Very big gains for very persistent outcomes such as tests scores
- Using OD
  - Pre-test score will be used as a covariate, r2 is it correlation over time.

The design factors that influence power

1. Clustered design
2. Availability of a Baseline
3. **Availability of Control Variables, and Stratification.**
4. The type of hypothesis that is being tested.

Stratified samples

- Stratification will reduce the sample size needed to achieve a given power (you saw this in real time in the Balsakhi exercise).
- The reason is that it will reduce the variance of the outcome of interest in each strata (and hence increase the standardized effect size for any given effect size)
- Example: if you randomize within school and grade which class is treated and which class is control:
  - The variance of test score goes down because age is controlled for

- Common stratification variables:
  - Baseline values of the outcomes when possible
  - We expect the treatment to vary in different subgroups

The design factors that influence power

1. Clustered design
2. Availability of a Baseline
3. Availability of Control Variables, and Stratification.
4. **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?)**

Conclusions

- Power calculations involve some guess work.
- They also involve some pilot testing before the proper experiment begins
- They can tell you:
  - How many treatments to have
  - How to trade off more clusters vs. more observations per cluster
  - Whether it's feasible or not
- It's critical to do as best you can; a study with low power likely wastes time and money