

EXERCISE C: HOW TO DO POWER CALCULATIONS IN OPTIMAL DESIGN SOFTWARE

TABLE OF CONTENTS

| | |
|--|----|
| Key Vocabulary..... | 1 |
| Introduction..... | 2 |
| Using the Optimal Design Software..... | 2 |
| Estimating Sample Size for a Simple Experiment..... | 9 |
| Some Wrinkles: Limited Resources and Imperfect Compliance..... | 12 |
| Clustered Designs..... | 14 |

KEY VOCABULARY

- 1. POWER:** The likelihood that, when a program/treatment has an effect, you will be able to distinguish the effect from zero i.e. from a situation where the program has no effect, given the sample size.
- 2. SIGNIFICANCE:** The likelihood that the measured effect did not occur by chance. Statistical tests are performed to determine whether one group (e.g. the experimental group) is different from another group (e.g. comparison group) on certain outcome indicators of interest (for instance, test scores in an education program.)
- 3. STANDARD DEVIATION:** For a particular indicator, a measure of the variation (or spread) of a sample or population. Mathematically, this is the square root of the variance.
- 4. STANDARDIZED EFFECT SIZE:** A standardized (or normalized) measure of the [expected] magnitude of the effect of a program. Mathematically, it is the difference between the treatment and control group (or between any two treatment arms) for a particular outcome, divided by the standard deviation of that outcome in the control (or comparison) group.
- 5. CLUSTER:** The unit of observation at which a sample size is randomized (e.g. school), each of which typically contains several units of observation that are measured (e.g. students). Generally, observations that are highly correlated with each other should be clustered and the estimated sample size required should be measured with an adjustment for clustering.

6. INTRA-CLUSTER CORRELATION COEFFICIENT (ICC): A measure of the correlation between observations within a cluster. For instance, if your experiment is clustered at the school level, the ICC would be the level of correlation in test scores for children in a given school relative to the overall correlation of students in all schools.

INTRODUCTION

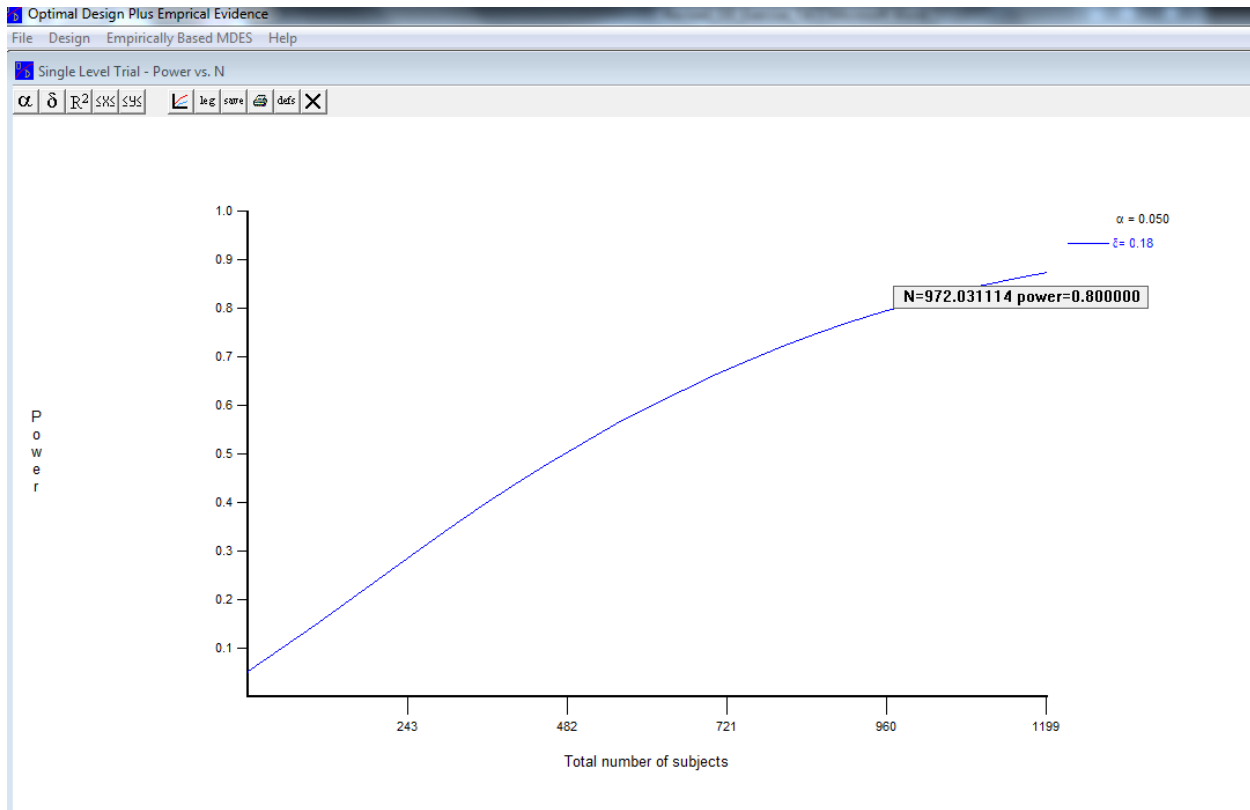
This exercise will help explain the trade-offs to power when designing a randomized trial. Should we sample every student in just a few schools? Should we sample a few students from many schools? How do we decide?

We will work through these questions by determining the sample size that allows us to detect a specific effect with at least 80 percent power, which is a commonly accepted level of power. Remember that power is the likelihood that when a program/treatment has an effect, you will be able to distinguish it from zero in your sample. Therefore at 80% power, if an intervention's impact is statistically significant at exactly the 5% level, then for a given sample size, we are 80% likely to detect an impact (i.e. we will be able to reject the null hypothesis.)

In going through this exercise, we will use the example of an education intervention that seeks to raise test scores. This exercise will demonstrate how the power of our sample changes with the number of school children, the number of children in each classroom, the expected magnitude of the change in test scores, and the extent to which children within a classroom behave more similarly than children across classrooms. We will use a software program called *Optimal Design*, developed by Stephen Raudenbush et al. with funding from the William T. Grant Foundation. Additional resources on research designs can be found on their web site.

USING THE OPTIMAL DESIGN SOFTWARE

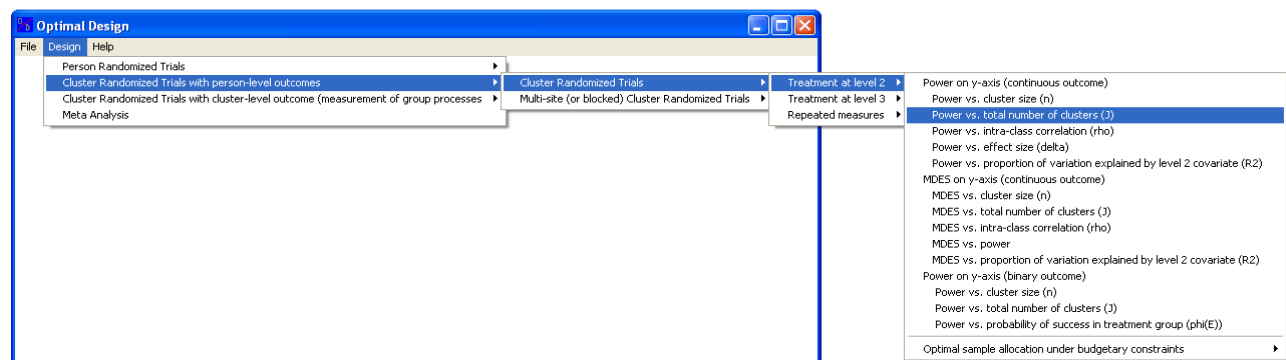
Optimal Design produces a graph that can show a number of comparisons: Power versus sample size (for a given effect), effect size versus sample size (for a given desired power), with many other options. The chart on the next page shows power on the y-axis and sample size on the x-axis. In this case, we inputted an effect size of 0.18 standard deviations (explained in the example that follows) and we see that we need a sample size of 972 to obtain a power of 80%.



We will now go through a short example demonstrating how the OD software can be used to perform power calculations. If you haven't downloaded a copy of the OD software yet, you can do so from the following website (where a software manual is also available):

http://sitemaker.umich.edu/group-based/optimal_design_software

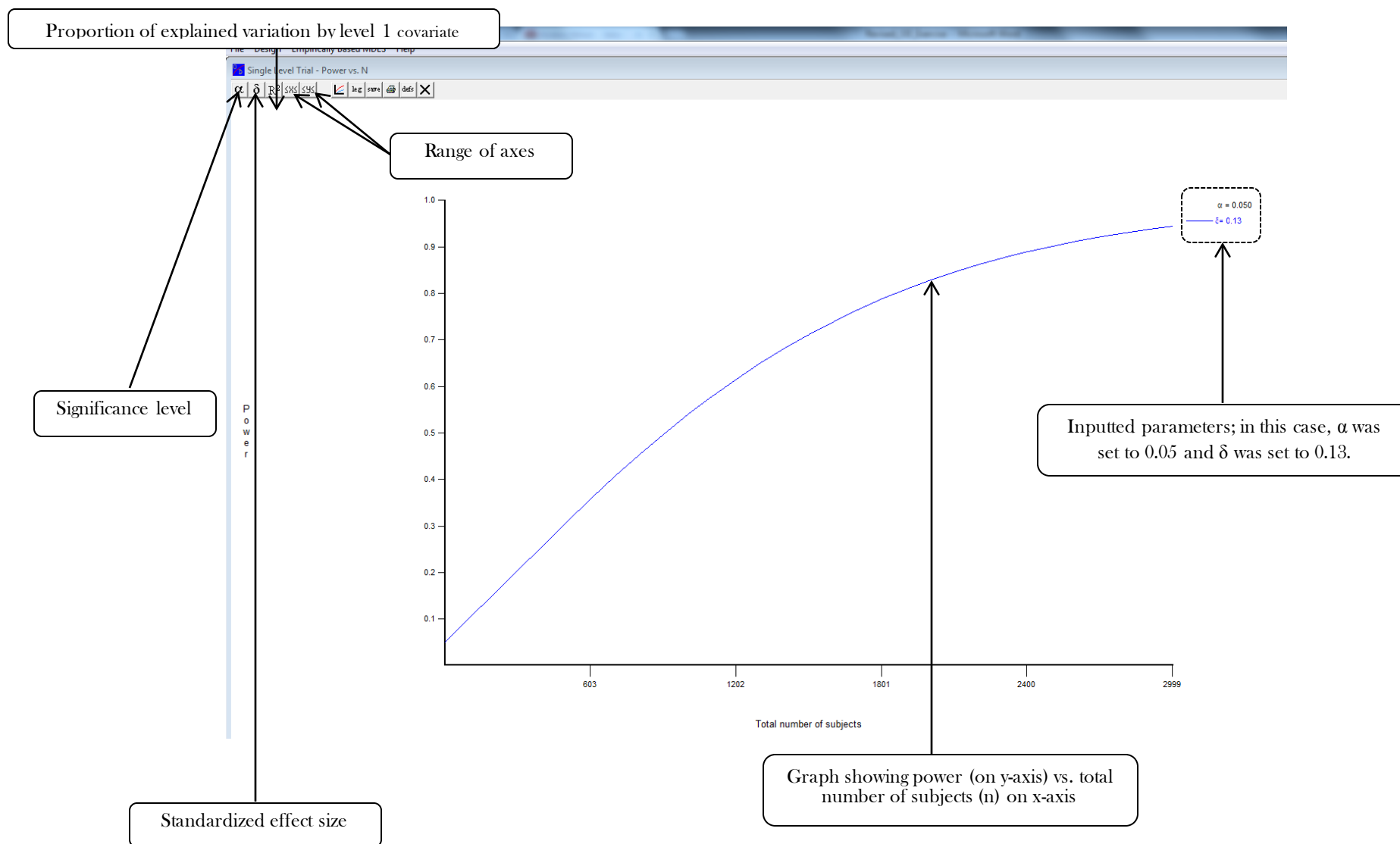
Running the HLM software file “od” should give you a screen which looks like the one below:



The various menu options under “Design” allow you to perform power calculations for randomized trials of various designs.

Let's work through an example that demonstrates how the sample size for a simple experiment can be calculated using OD. Follow the instructions along as you replicate the power calculations presented in this example, in OD. On the next page we have shown a sample OD graph, highlighting the various components that go into power calculations. These are:

- Significance level (α): For the significance level, typically denoted by α , the default value of 0.05 (i.e. a significance level of 95%) is commonly accepted.
- Standardized effect size (δ): Optimal Design (OD) requires that you input the standardized effect size, which is the effect size expressed in terms of a normal distribution with mean 0 and standard deviation 1. This will be explained in further detail below. The default value for δ is set to 0.200 in OD.
- Proportion of explained variation by level 1 covariate (R^2): This is the proportion of variation that you expect to be able to control for by including covariates (i.e. other explanatory variables other than the treatment) in your design or your specification. The default value for R^2 is set to 0 in OD.
- Range of axes ($\leq x \leq$ and $\leq y \leq$): Changing the values here allows you to view a larger range in the resulting graph, which you will use to determine power.



We will walk through each of these parameters below and the steps involved in doing a power calculation. Prior to that though, it is worth taking a step back to consider what one might call the “paradox of power”. Put simply, in order to perfectly calculate the sample size that your study will need, it is necessary to know a number of things: the effect of the program, the mean and standard deviation of your outcome indicator of interest for the control group, and a whole host of other factors that we deal with further on in the exercise. However, we cannot know or observe these final outcomes until we actually conduct the experiment! We are thus left with the following paradox: In order to conduct the experiment, we need to decide on a sample size...a decision that is contingent upon a number of outcomes that we cannot know without conducting the experiment in the first place.

It is in this regard that power calculations involve making careful assumptions about what the final outcomes are likely to be – for instance, what effect you realistically expect your program to have, or what you anticipate the average outcome for the control group being. These assumptions are often informed by real data: from previous studies of similar programs, pilot studies in your population of interest, etc. The main thing to note here is that to a certain extent, power calculations are more of an art than a science. However, making wrong assumptions will not affect accuracy (i.e. will not bias the results). It simply affects the precision with which you will be able to estimate your impact. Either way, it is useful to justify your assumptions, which requires carefully thinking through the details of your program and context.

With that said, let us work through the steps for a power calculation using an example. Say your research team is interested in looking at the impact of providing students a tutor. These tutors work with children in grades 2, 3 and 4 who are identified as falling behind their peers. Through a pilot survey, we know that the average test scores of students before receiving tutoring is 26 out of 100, with a standard deviation of 20. We are interested in evaluating whether tutoring can cause a 10 percent increase in test scores.

1) Let’s find out the minimum sample that you will need in order to be able to detect whether the tutoring program causes a 10 percent increase in test scores. Assume that you are randomizing at the school level i.e. there are treatment schools and control schools.

I. What will be the mean test score of members of the control group? What will the standard deviation be?

Answer: To get the mean and standard deviation of the control group, we use the mean and standard deviation from our pilot survey i.e. mean = 26 and standard deviation = 20. Since we do not know how the control group’s scores will change, we assume that the control group’s scores will not increase absent the tutoring program and will correspond to the scores from our pilot data.

II. If the intervention is supposed to increase test scores by 10%, what should you expect the mean and standard deviation of the treatment group to be after the intervention? Remember, in this case we are considering a 10% increase in scores over the scores of the control group, which we calculated in part I.

Answer: Given that the mean of the control group is 26, the mean with a 10% increase would be $26 \times 1.10 = \underline{28.6}$. With no information about the sample distribution of the treatment group after the intervention, we have no reason for thinking that there is a higher amount of variability within the treatment group than the control group (i.e. we assume homogeneous treatment impacts across the population). In reality, the treatment is likely to have heterogeneous i.e. differential impacts across the population, yielding a different standard deviation for the treatment group. For now, we assume the standard deviation of the treatment group to be the same as that of the control group i.e. 20.

- III. Optimal Design (OD) requires that you input the standardized effect size, which is the effect size expressed in terms of a normal distribution with mean 0 and standard deviation 1. Two of the most important ingredients in determining power are the *effect size* and the *variance* (or standard deviation). The standardized effect size basically combines these two ingredients into one number. The standardized effect size is typically denoted using the symbol δ (delta), and can be calculated using the following formula:

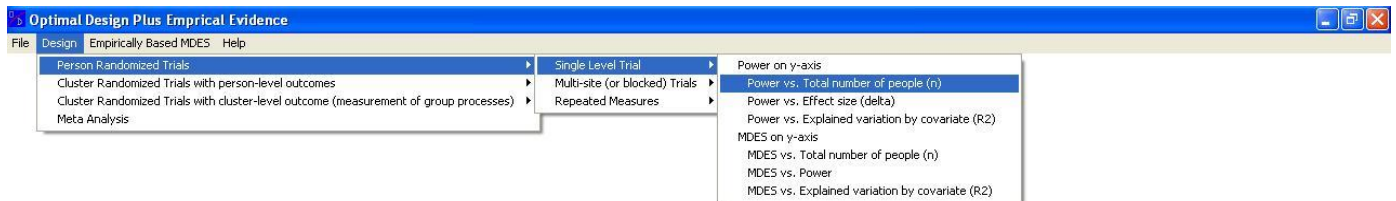
$$\delta = \frac{(\text{Treatment Mean} - \text{Control Mean})}{(\text{Standard Deviation})}$$

Using this formula, what is δ ?

Answer: $\delta = \frac{(28.6-26)}{20} = 0.13$

- IV. Now use OD to calculate the sample size that you need in order to detect a 10% increase in test scores. You can do this by navigating in OD as follows:

Design → Person Randomized Trials → Single Level Trial → Power vs. Total number of people (n)



There are various parameters that you will be asked to fill in:



You can do this by clicking on the button with the symbol of the parameter. To reiterate, the parameters are:

- Significance level (α): For the significance level, typically denoted by α , the default value of 0.05 (i.e. a significance level of 95%) is commonly accepted.
- Standardized effect size (δ): The default value for δ is set to 0.200 in OD. However, you will want to change this to the value that we computed for δ in part C.
- Proportion of explained variation by level 1 covariate (R^2): This is the proportion of variation that you expect to be able to control for by including covariates (i.e. other explanatory variables other than the treatment) in your design or your specification. We will leave this at the default value of 0 for now and return to it later on.
- Range of axes ($\leq x \leq$ and $\leq y \leq$): Changing the values here allows you to view a larger range in the resulting graph, which you will use to determine power; we will return to this later, but can leave them at the default values for now.

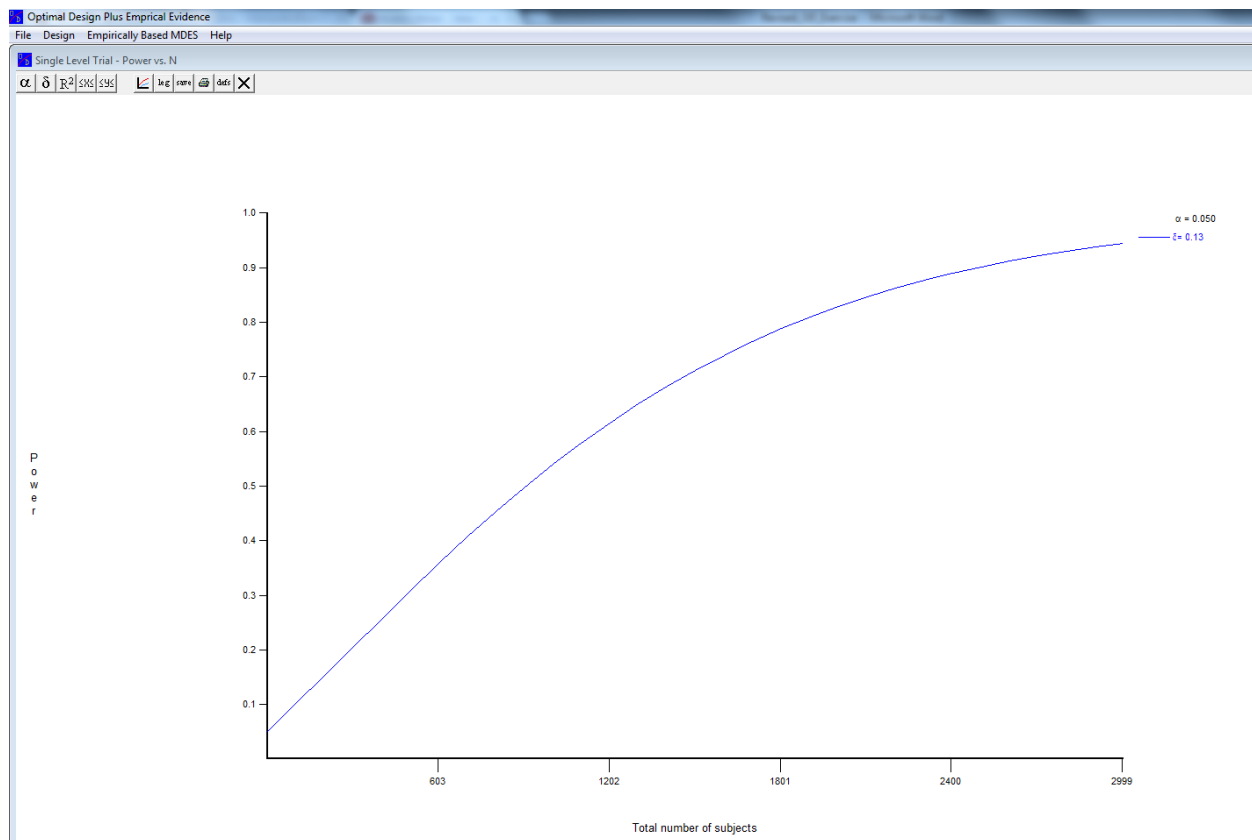
What will your total sample size need to be in order to detect a 10% increase in test scores at 80% power?

Answer: Once you input the various values above into the appropriate cells, you will get a plot with power on the y-axis and the total number of subjects on the x-axis. Click your mouse on the plot to see the power and sample size for any given point on the line.

Power of 80% (0.80 on the y-axis of your chart) is typically considered an acceptable threshold. This is the level of power that you should aim for while performing your power calculations. You will notice that just inputting the various values above does not allow you to see the number of subjects required for 80% power. You will thus need to increase the range of your x-axis; set the maximum value at 3000. This will yield a plot that looks the one on the following page.

To determine the sample size for a given level of power, click your mouse cursor on the graph line at the appropriate point. While this means that arriving at *exactly* a given level of power (say power of exactly 0.80) is difficult, a very good approximate (i.e. within a couple of decimal places) is sufficient for our purposes.

Clicking your mouse cursor on the line at the point where Power ~ 0.8 tells us that the total number of subjects, called “N”, is approximately **1,850**. OD assumes that the sample will be balanced between the treatment and control groups. **Thus, the treatment group will have $1850/2 = 925$ students and the control group will have $1850/2 = 925$ students as well.**



ESTIMATING SAMPLE SIZE FOR A SIMPLE EXPERIMENT

All right, now it is your turn! For the parts A – I below, leave the value of R^2 at the default of 0 whenever you use OD; we will experiment with changes in the R^2 value a little later.

You decide that you would like your study to be powered to measure an increase in test scores of 20% rather than 10%. Try going through the steps that we went through in the example above. Let's find out the minimum sample you will need in order to detect whether the tutoring program can increase test scores by 20%.

A. What is the mean test score for the control group? What is the standard deviation?

Mean:

Standard deviation:

B. If the intervention is supposed to increase test scores by 20%, what should you expect the mean and standard deviation of the treatment group to be after the intervention?

Mean:

Standard deviation:

C. What is the desired standardized effect size δ ? Remember, the formula for calculating δ is:

$$\delta = \frac{(\text{Treatment Mean} - \text{Control Mean})}{(\text{Standard Deviation})}$$

δ :

D. Now use OD to calculate the sample size that you need in order to detect a 20% increase in test scores.

Sample size (n):

Treatment:

Control:

- E. Is the *minimum* sample size required to detect a 10% increase in test scores larger or smaller than the minimum sample size required to detect a 20% increase in test scores? Intuitively, will you need larger or smaller samples to measure smaller effect sizes?

Answer:

- F. Your research team has been thrown into a state of confusion! While one prior study led you to believe that a 20% increase in test scores is possible, a recently published study suggests that a more conservative 10% increase is more plausible. What sample size should you pick for your study?

Answer:

- G. Both the studies mentioned in part F found that although average test scores increased after the tutoring intervention, the standard deviation of test scores also increased i.e. there was a larger spread of test scores across the treatment groups. To account for this, you posit that instead of 20, the standard deviation of test scores may now be 25 after the tutoring program. Calculate the new δ for an increase of 10% in test scores.

δ :

- H. For an effect of 10% on test scores, does the corresponding standardized effect size increase, decrease, or remain the same if the standard deviation is 25 versus 20? Without plugging the values into OD, all other things being equal, what impact does a higher standard deviation of your outcome of interest have on your required sample size?

Answer:

- I. Having gone through the intuition, now use OD to calculate the sample size required in order to detect a 10% increase in test scores, if the pre-intervention mean test scores are 26, with a standard deviation of 25.

Sample size (n):

Treatment:

Control:

- J. One way by which you can increase your power is to include covariates i.e. control variables that you expect will explain some part of the variation in your outcome of interest. For instance, baseline, pre-intervention test scores may be a strong predictor of a child's post-intervention test scores; including baseline test scores in your eventual regression specification would help you to isolate the variation in test scores attributable to the tutoring intervention more precisely. You can account for the presence of covariates in your power calculations using the R^2 parameter, in which you specify what proportion of the eventual variation in your outcome of interest is attributable to your treatment condition.

Say that you have access to the pre-intervention test scores of children in your sample for the tutoring study. Moreover, you expect that pre-intervention test scores explain 50% of the variation in post-intervention scores. What size sample will you require in order to measure an increase in test scores of 10%, assuming standard deviation in test scores of 25, with a pre-intervention mean of 26. Is this more or less than the sample size that you calculated in part I?

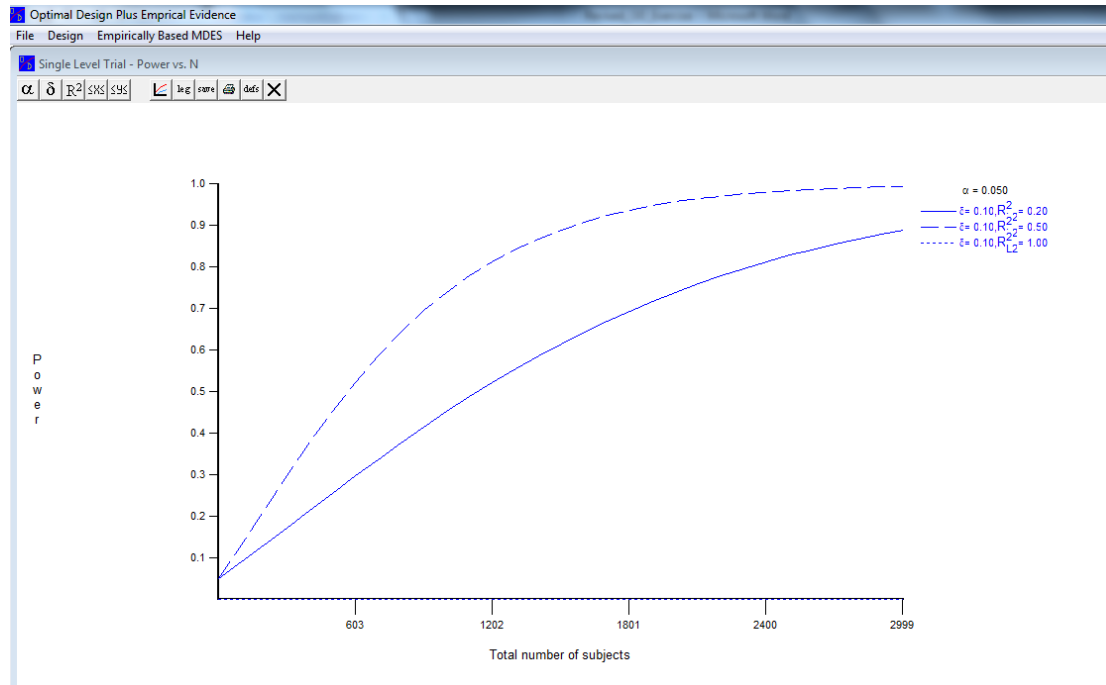
Sample size (n):

Treatment:

Control:

- K. One of your colleagues on the research team thinks that 50% may be too ambitious an estimate of how much of the variation in test scores post-intervention is attributable to baseline scores. She suggests that 20% may be a better estimate. What happens to your required sample size when you run the calculations from part J with an R^2 of 0.200 instead of 0.500? What happens if you set R^2 to be 1.000?

Tip: You can enter up to 3 separate values on the same graph for the R^2 in OD; if you do, you will end up with a figure like the one below:



Answer:

SOME WRINKLES: LIMITED RESOURCES AND IMPERFECT COMPLIANCE

- L. You find out that you only have enough funds to survey 1,200 children. Assume that you do not have data on baseline covariates, but know that pre-intervention test scores were 26 on average, with a standard deviation of 20. What standardized effect size (δ) would you need to observe in order to survey a maximum of 1,200 children and still retain 80% power? Assume that the R^2 is 0 for this exercise since you have no baseline covariate data.

Hint: You will need to plot "Power vs. Effect size (delta)" in OD, setting "N" to 1,200. You can do this by navigating in OD as follows: Design → Person Randomized Trials → Single Level Trial → Power vs. Effect Size (delta). Then, click on the point of your graph that roughly corresponds to power = 0.80 on the y-axis.

$\delta =$

- M. Your research team estimates that you will not realistically see more than a 10% increase in test scores due to the intervention. Given this information, is it worth carrying out the study on just 1,200 children if you are adamant about still being powered at 80%?

Answer:

- N. Your research team is hit with a crisis: You are told that you cannot force people to use the tutors! After some small focus groups, you estimate that only 40% of schoolchildren would be interested in the tutoring services. You realize that this intervention would only work for a very limited number of schoolchildren. You do not know in advance whether students are likely to take up the tutoring service or not. How does this affect your power calculations?

Answer:

- O. You have to “adjust” the effect size you want to detect by the proportion of individuals that actually gets treated. Based on this, what will be your “adjusted” effect size and the adjusted standardized effect size (δ) if you originally wanted to measure a 10% increase in test scores? Assume that your pre-intervention mean test score is 26, with a standard deviation of 20, you do not have any data on covariates, and that you can survey as many children as you want.

Hint: Keep in mind that we are calculating the average treatment effect for the entire group here. Thus, the lower the number of children that actually receives the tutoring intervention, the lower will be the measured effect size.

Answer:

- P. What sample size will you need in order to measure the effect size that you calculated in part O with 80% power? Is this sample bigger or smaller than the sample required when you assume that 100% of children take up the tutoring intervention (as we did in the example at the start)?

Sample size (n):

Treatment:

Control:

CLUSTERED DESIGNS

Thus far we have considered a simple design where we randomize at the *individual-level* i.e. school children are either assigned to the treatment (tutoring) or control (no tutoring) condition. However, spillovers could be a major concern with such a design: If treatment and control students are in the same *school*, let alone the same classroom, students receiving tutoring may affect the outcomes for students not receiving tutoring (through peer learning effects) and vice versa. This would lead us to get a biased estimate of the impact of the tutoring program.

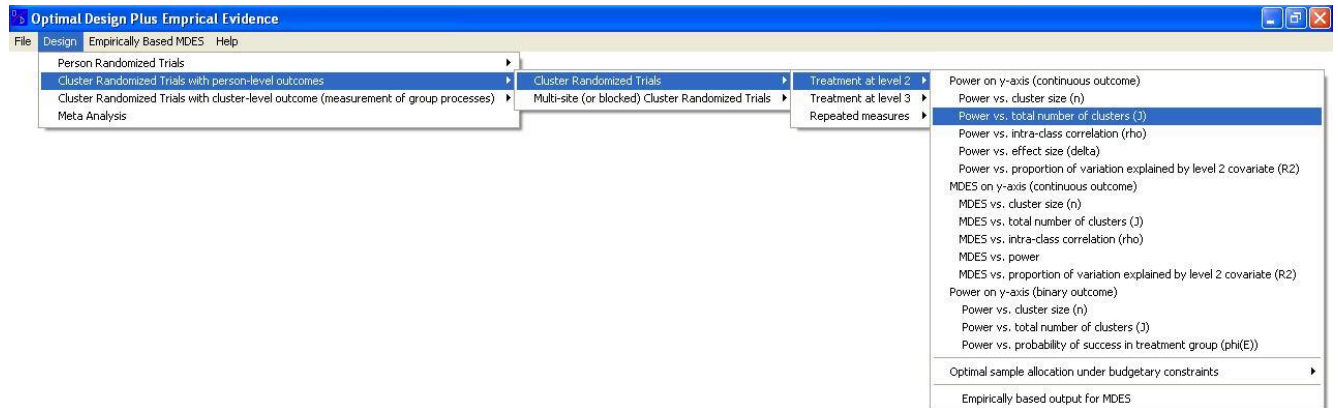
In order to preclude this, your research team decides that it would like to run a cluster randomized trial, randomizing at the *school-level* instead of the individual-level. In this case, each school forms a “cluster”, with all the students in a given school assigned to either the treatment condition, or the control one. Under such a design, the only spillovers that may show up would be across schools, a far less likely possibility than spillovers within schools.

Since the behavior of individuals in a given cluster will be correlated, we need to take an **intra-cluster or intra-class correlation (denoted by the Greek symbol ρ)** into account for each outcome variable of interest. Remember, ρ is a measure of the correlation between children within a given school (see key vocabulary at the start of this exercise.) ρ tells us how strongly the outcomes are correlated for units within the same cluster. If students from the same school were clones (no variation) and all scored the same on the test, then ρ would equal 1. If, on the other hand, students from the same schools were in fact independent and there was zero difference between schools or any other factor that affected those students, then ρ would equal 0.

The ρ or ICC of a given variable is typically determined by looking at pilot or baseline data for your population of interest. Should you not have the data, another way of estimating the ρ is to look at other studies examining similar outcomes amongst similar populations. Given the inherent uncertainty with this, it is useful to consider a range of ρ s when conducting your power calculations (a sensitivity analysis) to see how sensitive they are to changes in ρ . We will look at this a little further on. While the ρ can vary widely depending on what you are looking at, values of less than 0.05 are typically considered low, values between 0.05-0.20 are considered to be of moderate size, and values above 0.20 are considered fairly high. Again, what counts as a low ρ and what counts as a high ρ can vary dramatically by context and outcome of interest, but these ranges can serve as initial rules of thumb.

Based on a pilot study and earlier tutoring interventions, your research team has determined that the ρ is 0.17. You need to calculate the total sample size to measure a 15% increase in test scores (assuming that test scores at the baseline are 26 on average, with a standard deviation of 20, setting R^2 to 0 for now). You can do this by navigating in OD as follows:

Design → Cluster Randomized Trials with person-level outcomes → Cluster Randomized Trials → Treatment at Level 2 → Power vs. total number of clusters (J)



In the bar at the top, you will see the same parameters as before, with an additional option for the intra-cluster correlation. Note that OD uses “n” to denote the cluster size here, not the total sample size. OD assigns two default values for the effect size (δ) and the intra-cluster correlation (ρ), so do not be alarmed if you see four lines on the chart. Simply delete the default values and replace them with the values for the effect size and intra-cluster correlation that you are using.

Q. What is the effect size (δ) that you want to detect here? Remember that the formula for calculating δ is:

$$\delta = \frac{(\text{Treatment Mean} - \text{Control Mean})}{(\text{Standard Deviation})}$$

δ :

R. Assuming there are 40 children per school, how many schools would you need in your clustered randomized trial?

Answer:

S. Given your answer above, what will the total size of your sample be?

Sample size:

Treatment:

Control:

T. What would the number of schools and total sample size be if you assumed that 20 children from each school were part of the sample? What about if 100 children from each school were part of the sample?

| | 20 children per school | 40 children per school | 100 children per school |
|------------------------|------------------------|------------------------|-------------------------|
| Number of schools: | | 160 | |
| Total no. of students: | | 6,400 | |

- U. As the number of clusters increases, does the total number of students required for your study increase or decrease? Why do you suspect this is the case? What happens as the number of children per school increases?

Answer:

- V. You realize that you had read the pilot data wrong: It turns out that the ρ is actually 0.07 and not 0.17. Now what would the number of schools and total sample size be if you assumed that 20 children from each school were part of the sample? What about if 40 or 100 children from each school were part of the sample?

| | 20 children per school | 40 children per school | 100 children per school |
|------------------------|------------------------|------------------------|-------------------------|
| Number of schools: | | | |
| Total no. of students: | | | |

- W. How does the total sample size change as you increase the number of individuals per cluster in part V? How do your answers here compare to your answers in part T?

Answer:

- X. Given a choice between offering the tutors to more children in each school (i.e. adding more individuals to the cluster) versus offering tutors in more schools (i.e. adding more clusters), which option is best *purely from the perspective of improving statistical power*? Can you imagine a situation when there will not be much difference between the two from the perspective of power?

Answer: