J-PAL
ABDUL LATIF JAMEEL POVERTY ACTION LAB

EXERCISE: HOW TO DO POWER CALCULATIONS

Table of Contents

KEY VOCABULARY	
Significance level	The probability of committing a type I error (a false positive: concluding that the program has an effect when it actually does not). Statistical tests are typically performed at significance levels of 1%, 5%, or sometimes 10% to determine whether one group (e.g., the experimental group) is different from another group (e.g., the comparison group) on certain outcome indicators of interest (for instance, test scores in an education program). The significance level is typically denoted by alpha (α).
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.
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.
Type II error	A false negative: finding no evidence of impact when a program/treatment actually has an effect. Type II error is often denoted by kappa (κ).
Power	The likelihood of avoiding a type II error, that is, the probability that your statistical test will distinguish the program effect (correctly) from zero when the program/treatment actually has an effect, given the sample size and the population the sample is drawn from). Power mirrors the significance level, α: as α increases (e.g., from 1% to 5%), the probability of rejecting the null hypothesis increases, which translates to a more powerful test.
Cluster	The unit level at which a sample is randomized (e.g., school), each of which typically contains several units of observation that are measured (e.g., students). Generally, observations within the same unit of randomization that are potentially correlated with each other should be clustered, and the required sample size should be calculated with an adjustment for clustering.
Intra-cluster correlation coefficient (ICC)	A measure of the correlation between observations within the same cluster. For instance, if your experiment is clustered at the school level and the outcome of interest is test scores, the ICC would be the level of correlation in test scores for children in a given school relative to the overall distribution of test scores of students in all schools.

INTRODUCTION

In this exercise, we will practice power calculations using estimates of effect sizes and outcome variance. The exercise will also help explain the considerations that go into determining sample size and randomization design when designing a randomized evaluation for an education intervention. Should the intervention be at the school level or the student level? Should we sample every student in just a few schools? Should we sample a few students from many schools? How many students or schools must we sample to confidently avoid making a type II error?

We will work through these questions by determining the sample size that allows us to detect a specific effect with at least 80% power, which is a commonly accepted level of power (such as by organizations that fund research). Recall that power is the probability of avoiding a false negative, also known as a type II error. That is, power is the likelihood that, when a treatment or program has an effect, you will be able to distinguish this effect from zero in your sample. Therefore, if our sample is chosen for 80% power and we accept that an intervention had an impact if it is statistically significant at the 5% level (a commonly accepted level of significance), then at the given sample size, we are 80% likely to correctly reject the null hypothesis (typically, the null hypothesis is that the program had no effect). Alternatively, if a study has not been set up to achieve a commonly accepted level of power, it is considered to be "underpowered" and is at risk of a type II error.

Throughout this exercise, we will use the example of an education intervention that seeks to raise test scores. We will explore how our study's power changes with the total number of students, the number of students in each classroom, the expected magnitude of the change in test scores, and the extent to which students within a classroom appear more similar than students across classrooms.

We will walk through how to do simple power calculations in two ways: (1) by using results from a similar study to estimate the effect size and standard deviation of the outcome variable for our program, and (2) by using pilot data from our program (the Balsakhi study in the lecture) to calculate these components ourselves.[1]

USING THE EGAP POWER CALCULATOR

For this exercise, we will use a calculator for power calculations developed by Alexander Coppock for EGAP (Evidence in Governance and Politics). The calculator was developed using the Shiny package in R and can be used to conduct power calculations for individual-level randomization, clustered designs, and with binary or continuous outcome variables. The calculator can be accessed at https://egap.shinyapps.io/power-app/.

ESTIMATING SAMPLE SIZE USING RESULTS FROM A SIMILAR STUDY

Let's work through this part together. Recall, the key components needed to estimate sample size using a simple (non-clustered) design are:

[1] The "similar study" results shown in Table 1 are fictional and were created for the purposes of this exercise. The "pilot data" provided is a random subset of the actual Balsakhi program survey data. The full dataset for the Balsakhi program is available on the J-PAL Dataverse: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/UV7ERB.

Significance level (α): Typically denoted by α, the significance level is the probability of a type I error, or falsely concluding that there is an effect when there is none (falsely rejecting the null hypothesis). The EGAP calculator default value of α=0.05 is commonly accepted.

- **Power (1-κ):** The probability of avoiding a type II error, where a type II error is falsely concluding there is no effect when there is one. Power is typically set at 80%, though some researchers aim for 90%.
- The **effect size** of the program for the outcome of interest.
- The **variance** (or the root of the variance, the **standard deviation**) of the outcome of interest.

More complex designs, such as clustered designs or using covariates to improve the precision of your estimates, require additional components to estimate sample size, which will be explained in further detail below.

The paradox of power is that we cannot know two of the above components, namely the effect size and the variance of the outcome of interest, until we conduct the experiment! That is, in order to conduct the experiment, we need to decide on a sample size—but this decision is contingent on a number of outcomes that we cannot know without conducting the experiment in the first place.

In this regard, power calculations involve making careful assumptions about certain outcomes, such as the effect you realistically expect your program may have or the variation you expect in the outcome variable. These assumptions are often informed by real data, such as from previous studies of similar programs, or pilot studies in your population of interest. Making wrong assumptions will not bias the results of the study but will affect the likelihood of a type II error, or failing to detect an effect when there is one. Regardless of the source of the data you use to inform your power calculations, it is important to justify your assumptions, which requires carefully thinking through the details of your program and context.

We will start by using data from a previous study looking at a similar program to inform our power calculations. Table 1 shows the regression results from a program in Andhra Pradesh, India that sought to increase student test scores through intensive tutoring. The table also shows the mean and standard deviation of the pre-treatment outcome variable, pre-test scores. We will use the effect size and distribution of test scores in this study as a benchmark to conduct power calculations for our own study.

Table 1

Effect of the Andhra Pradesh Tutoring Program on Post-test Scores	
Received tutor	3.8*** (1.12)
Constant	35.9 (0.83)
Average pre-test score	36.4

Standard deviation of pre-test score	15.2
N	694

Notes: Test scores are out of 60 possible points. Standard error is in parentheses.
***Statistically significant at the 1% level.

A. What is β, the treatment effect size?

B. What is the standard deviation of the dependent variable?

If you haven't already, open the EGAP calculator in your web browser by going to: https://egap.shinyapps.io/power-app/. Next, set the desired power and significance – for this exercise we'll use the standard values of α=0.05 and 1-κ=0.8. Keep the maximum number of subjects at the default and leave the "Clustered design?" and "Binary dependent variable?" boxes unchecked right now; we'll return to these later.

C. Plug in the values you found in questions A and B. Given these parameters and a significance level of α=0.05, what is the sample size needed for 80% power?

D. The EGAP calculator gives the total sample size. Assuming half allocated to treatment and half allocated to control, how many do you need in your treatment and control groups?

Number in treatment:
Number in control:

ESTIMATING SAMPLE SIZE USING DATA FROM A PILOT STUDY

Now it's your turn! Suppose you have data from a pilot study your team did for this project. This data can be found in (xls file). Here, test scores are again our outcome of interest, and the pilot study was done in the same population from which you'll draw your sample for the main study of the Balsakhi tutoring program.

E. What is the mean and standard deviation of the dependent variable?

Mean:
Standard deviation:

F. After the tutoring program, what do you expect for the mean test score in the control group? What standard deviation do you expect?

Mean:

Standard deviation:

After deliberations with your partner organization, you've decided that you need an effect size of at least 10% for the program to be worth the its costs, which is roughly what the previous study from **A-D** found.

G. If you observe a 10% increase in test scores as a result of the tutoring program, what is the mean test score for the treatment group after the intervention? What do you expect for the standard deviation of test scores in the treatment group? What is the effect size?

Mean test score in treatment group:

Standard deviation of test scores in treatment group:

Effect size:

H. Given $\alpha=0.05$ and the standard deviation of the outcome variable you found above, what is the minimum sample size you need to detect the effect size you found in **G** with 80% power?

Total number of participants:
Number in treatment:
Number in control:

A new study has come out finding a 5% increase in test scores as a result of a tutoring program in a different city in Gujarat (the same state where your program will take place). While the prior study of the tutoring program in Andhra Pradesh led you to believe that a 10% increase in test scores is possible, the more recent study suggests that a smaller increase of 5% is more reasonable.

I. What is the effect size for a 5% increase in test scores?

J. Now calculate the minimum sample size that is needed to detect the effect size you found in **I**. Use the standard deviation you found in part **G.**

Total number of students:
Number in treatment:
Number in control:

K. Explain what you found about sample sizes needed to detect a 5% increase in test scores versus a 10% increase in test scores, given no changes in the variance of the outcome variable. Intuitively, will you need larger or smaller samples to measure effect sizes that are smaller, relative to their spread? Why?

L. Recall that in the first part of the exercise, you used an effect size of 10%, based on results from a previous study, to do your power calculations. With the new study suggesting that a 5% increase in test scores is more reasonable, you are now faced with a dilemma: what sample size should you pick for your study, and why?

While the new study found that test scores increased after the tutoring intervention, it also found that the standard deviation of test scores increased, meaning that there was a larger spread of test scores across the treatment group. This is because students respond differently to tutoring; some students' test scores increased dramatically after the tutoring program, while others' increased only slightly. To account for the higher variance in test scores, you posit that instead of the standard deviation you found in **G**, the standard deviation of test scores may now be 16.5 after the tutoring program.

M. Without going through the calculations, does the minimum sample size needed to detect a 10% increase in test scores increase, decrease, or remain the same when the standard deviation of test scores rises to 16.5?

N. Having gone through the intuition, now calculate the minimum sample size needed to detect a 10% increase in test scores, given a standard deviation in test scores of 16.5.

LIMITED RESOURCES AND IMPERFECT COMPLIANCE

Sometimes, rather than calculate a budget based on sample size, we have a maximum budget and need to decide whether it is worth doing the study (that is, whether we are sufficiently powered to detect a given effect size, conditional on budgetary limitations).

O. You find out that you only have enough funds for a sample size of 2400 in total. Using the more recent paper's estimate of a roughly 5% increase in test scores and a standard deviation of test scores of 16.5, what is the power of your experiment? (An approximate answer is okay; it's hard to get exact power on the calculator this way.) Is it worth carrying out the study on just 2400 students? How would you determine this?

P. If you use the 10% increase in test scores as suggested by the first study, is it worth carrying out the study on just 2400 students? What is the power of your experiment? Assume a standard deviation of test scores of 16.5.

Q. Your research team has conducted some focus groups and determined that only 40% of students would be interested in the tutoring services—that is, not every student who is offered the tutoring program will choose to attend. How does this affect your power calculations? Will the required sample size to detect a 10% increase in test scores increase, decrease, or remain the same? Why? Here, you should assume that noncompliance is random—otherwise, and as discussed in case study 4, we would have to worry about selection bias from noncompliance (for example, weaker students may be more likely to take up the tutoring services, which would give you an underestimate of the program's impact).

CLUSTERED DESIGNS

Thus far we have considered a simple design where we randomize at the individual level, where students 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 (such as through peer learning effects) and vice versa. This would lead us to get a **biased** estimate of the impact of the tutoring program. Here, we would likely underestimate the effect of the tutoring program, because students in the control group would benefit from it as well.

To avoid this issue, your research team decides 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 of the students in a given school assigned to either the treatment group or the comparison group. Under such a design, the only spillovers that may show up would be across schools, a far less likely possibility than spillovers within schools.

Generally, individuals within a cluster tend to be more similar to each other than individuals from different clusters. For example, students in the same school share the same teachers, the same peers, and may share similarities in socioeconomic status and other factors that help determine school performance. This correlation in behavior of individuals within a given cluster is called **intra-cluster** or **intra-class correlation**, and we need to account for it in our power calculations using rho **(ρ)**, the **intra-cluster correlation coefficient (ICC)**, for our outcome of interest. Remember, ρ is a measure of the degree of similarity between children within a given school (see key vocabulary at the start of this exercise); it tells us how strongly the outcomes are correlated for units within the same cluster (specifically, it is the share of the variance *between* clusters relative to the overall variance). If students from the same school all scored exactly the same on the test, then ρ would equal 1. If, on the other hand, test scores of students from the same school are independent, and there was zero correlation in test scores of students within the same school, then ρ would equal 0. Realistically, ρ will fall somewhere between these two extremes.

The ICC (ρ) of a given variable is typically determined by looking at pilot or baseline data for your population of interest. Should you not have this data, another way of estimating ρ is to look at other studies examining similar outcomes amongst similar populations. Given the inherent uncertainty with extrapolating across contexts and populations, it is useful to consider a range of ρs when conducting your power calculations to see how sensitive they are to changes in ρ, a process known as sensitivity analysis. We will look at this a little further on; assumptions about ρ will have important implications for power calculations. While 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 versus high ρ can vary dramatically by context and outcome of interest, but these ranges serve as initial rules of thumb.

First, let's look at how power changes with the ICC. Start by checking the "Clustered Design?" box in the EGAP calculator. With this box checked, the blue line on the graph shows power with individual-level randomization, while the green line shows power for a clustered design, both for a given a set of parameters (significance level, treatment effect size, standard deviation of the outcome variable, ICC, number of clusters per arm, power target, and sample size). Keep the significance level at α=0.05 and the power target at 0.8, and set the treatment effect size based on the 10% increase in test scores you found earlier, with a standard deviation of 16.5. Keep the number of clusters per arm at 40, but increase the maximum number of subjects to 5000 (which rescales the graph and lets you see more easily what's going on).

R. The ICC default setting on this calculator is 0.5. What happens to the green line, relative to the blue line, as you slowly increase it to 1? What happens as you slowly decrease it down to 0? Intuitively, why is the green line moving closer to or further from the blue line as you change the ICC?

Based on the pilot study and earlier tutoring interventions, your research team has estimated a ρ of 0.11. You need to calculate the total sample size to measure a 10% increase in test scores (assuming that the mean test score at baseline is 33, with a standard deviation of 16.5). You can do this by checking the clustered design box in the EGAP calculator and adjusting the intra-cluster correlation bar to 0.11.

S. Change the number of clusters per arm to 55. Given 55 clusters per arm (so 110 clusters in total), a 10% increase in test scores, a standard deviation of 16.5, and an ICC of 0.11, how many subjects in total do you need for 80% power? How many in each cluster? How does this compare to the sample size you found in part **N,** using individual-level randomization? Why?

With an individual-level randomization, we can only manipulate the number of participants in the study. But with a clustered design, we can manipulate the number of clusters and the number of participants in each cluster. The two affect power in slightly different ways and have different costs—it is typically going to be cheaper to add participants to a cluster than to add clusters, though sometimes there is a limit on the number of participants in a cluster (e.g., a set number of students in a class). We'll now examine how manipulating the number of clusters versus the number of participants per cluster affects power, starting with the number of clusters.

T. Using the same effect size, standard deviation, and ICC as in **S**, how many students do you need in total if you have 60 clusters per arm, or 120 clusters in total? How many students per cluster is this? Fill in the table below.

U. If you have 50 clusters per arm, or 100 clusters in total, how many students do you need per cluster and in total? Fill in the table below.

	50 schools per arm (100 schools in total)	55 schools per arm (110 schools in total)	60 schools per arm (120 schools in total)
Number of students per school:			
Total no. of students:			

V. 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?

W. You realize that you had read the pilot data wrong: It turns out that ρ is actually 0.07 and not 0.11. Now what would the number of students per cluster and total number of students if you had 50 schools per arm (or 100 schools in total)? What about with 55 schools per arm or 60 schools per arm? Fill in the table below.

	50 schools per arm (100 schools in total)	55 schools per arm (110 schools in total)	60 schools per arm (120 schools in total)
Number of students per school:			
Total no. of students:			

X. How do your answers here compare to your answers in part **U**? Why?

Y. 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?* What about from a cost perspective?

RESOURCES

3ie, “Power calculation for causal inference in social science: sample size and minimum detectable effect determination” (report and calculator tool: http://www.3ieimpact.org/evidence-hub/publications/working-papers/power-calculation-causal-inference-social-science-sample

Optimal Design program for power calculations: http://hlmsoft.net/od/

Institute for Fiscal Studies, “Going beyond simple sample size calculations: a practitioner’s guide for power calculations” (includes sample Stata code):
https://www.ifs.org.uk/uploads/publications/wps/WP201517_update_Sep15.pdf

http://blogs.worldbank.org/impactevaluations/power-calculations-101-dealing-with-incomplete-take-up

“Remedying Education: Evidence from Two Randomized Experiments in India” (background on the Balsakhi tutoring program): https://www.povertyactionlab.org/sites/default/files/publications/6%20Computer-Assisted%20Learning%20Project%20with%20Pratham%20in%20India%2007.pdf