J-PAL
ABDUL LATIF JAMEEL POVERTY ACTION LAB

Exercise: How to do Power Calculations

Table of Contents

KEY VOCABULARY	
Hypothesis	A proposed explanation for the effects of a given intervention. We can think of this as a claim to be tested. Hypotheses are intended to be made prior to the implementation of the intervention. E.g., *Giving textbooks to students will improve student learning*.
Type I error / false positive	Falsely concluding that the program/treatment had an effect when it actually did not.
Significance level (α)	The maximal probability of obtaining a false positive we want to allow. Statistical tests are typically performed at significance levels of 1%, 5%, or sometimes 10% to determine whether one group (e.g., the treatment 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).
Type II error / false negative	Falsely concluding that the program/treatment had no effect when it actually did have an effect.
Power	The likelihood of avoiding a false negative. This is the probability that your statistical test will distinguish the program effect (correctly) from zero when the program/treatment actually has an effect.
Sample size (N)	The total number of units in the study, in both treatment and comparison groups. In a clustered randomization, the sample size is the total number of units across clusters.
Sample split	The proportion of the sample assigned to the treatment group. For a given sample size, power is maximized with an equal split between treatment arms.
Minimum detectable effect (MDE)	The effect size is the difference between the average of the outcome of interest in the treatment and control groups. The MDE is the minimum effect size that can be detected with a given level of statistical power, statistical significance, and sample size.
Standard deviation	This is a measure of the spread of a sample or population for a particular indicator. Mathematically, the standard deviation is the square root of the variance.
Unit of randomization	The level of observation (e.g., individual, household, school, village) at which treatment and comparison groups are randomly assigned.
Cluster	If the unit of observation is not the same as the unit of randomization (e.g., if you want to measure the impact of extended recess on student learning but randomize at the classroom level), this would be called cluster randomization. Groups with multiple observational units that will be randomized together (e.g., classrooms) are referred to as "clusters." Clustering should be used to reduce spillovers, but may also be necessary if

	assigning the treatment at the observational-unit level is not feasible or practical.
Intra-cluster correlation coefficient (ICC)	The ICC describes the degree of similarity between units within clusters. 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. The ICC is often denoted by rho (ρ).

Introduction

In this exercise, you will **practice doing power calculations** with given values for the different components of power. You will also get practical experience with how different RCT design choices affect the power of a study.

Throughout the exercise, we will use the example of a tutoring intervention that seeks to raise test scores for students in elementary schools. We will explore how our study's power changes with:

- The choice of randomizing at the student vs. the school level
- The total number of students, and the number of students in each school
- The expected magnitude of the change in test scores
- The extent to which students within a school appear more similar than students across schools

Imagine that you want to evaluate the tutoring program with an RCT, and you need to decide on the RCT design, including the sample size and at what level to randomize. We will refer to data from hypothetical studies that evaluate the impacts of similar programs to inform our power calculations.

Using the EGAP Power Calculator

For this exercise, we will use an **online power calculator** developed by Alexander Coppock for EGAP (Evidence in Governance and Politics).[1] The calculator can conduct power calculations for simple designs with individual-level or clustered randomization, and binary or continuous outcome variables. The calculator can be accessed at https://egap.shinyapps.io/power-app/. In its simplest form, the EGAP power

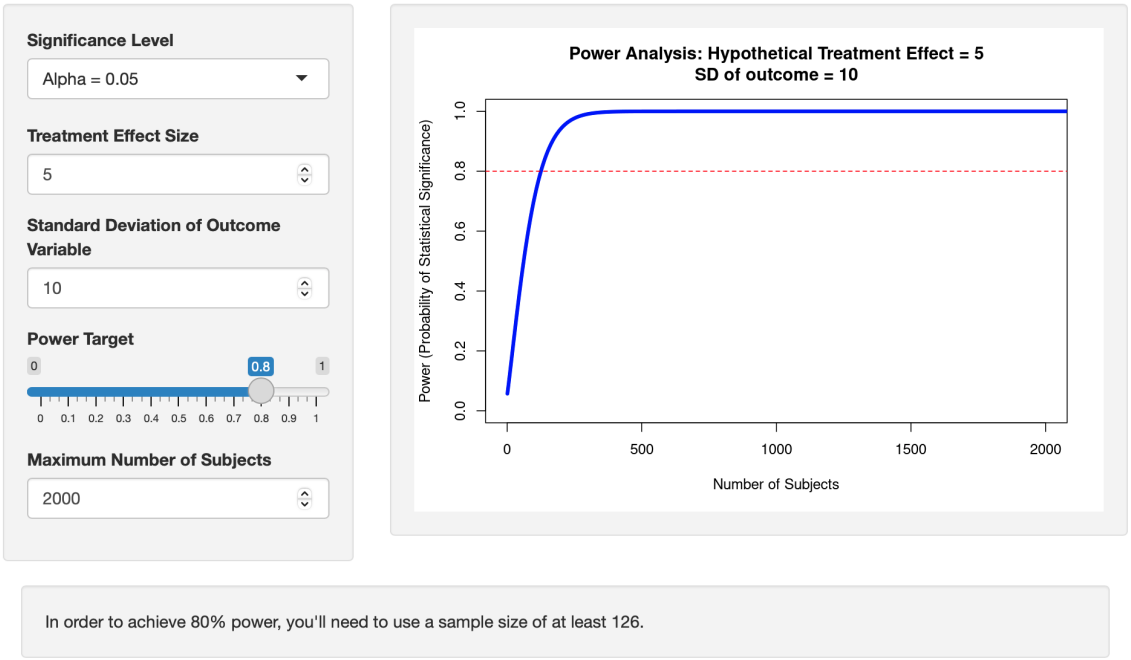
[1] While the EGAP calculator offers a useful tool to illustrate power calculations, researchers will typically use statistical programs such as Stata or R to conduct power calculations in practice. The calculator was developed using the Shiny package in R, but no knowledge of R is required to access or use the calculator.

calculator calculates the **required sample size** for an individually randomized study with one treatment arm, and a 50/50 sample split between the treatment and control groups. To calculate the required sample size for this simple design, you need to enter the following components:

- Significance Level (default is alpha = 0.05)
- Treatment Effect Size
- Standard Deviation of Outcome Variable
- Power Target (default is 80%)

The paradox of power calculations is that we don't know two important components of power—effect size and the standard deviation of the outcome of interest—until after we conduct the experiment. However, in order to conduct the experiment, we need to decide on a sample size, which requires conducting power calculations.

Therefore, power calculations involve making careful assumptions about certain parameters, 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 previous studies of similar programs, or data from pilot studies in your population of interest.



The graph on the power calculator shows how the power depends on the number of subjects in the RCT for a given significance level, effect size, and variance. The required number of subjects to reach the power target is given by the point at which the curve intersects with the power target, and displayed at the bottom of the page, e.g., "In order to achieve 80% power, you'll need to use a sample size of at least 126."

Note that the field for Maximum Number of Subjects only affects the horizontal axis on the graph—its value does not enter into the calculations.

You can ignore the code on the right-hand side of the screen (or under the graph, depending on your screen set-up). This is the R script behind the calculations, but you will not need to interact with that.

Estimating Sample Size Using Results From a Similar Study

We will start by using data from a previous study looking at a similar program to inform our power calculations. An RCT of a program that sought to improve student test scores through intensive tutoring in Andhra Pradesh, India, found that tutoring increased test scores by 3.8 points (out of 60 possible points). Before the tutoring intervention, student test scores had a mean of 36.4 points, and a standard deviation of 15.2 points.

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 significance and power – 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 for now. We'll return to these later.

A. Plug in the treatment effect size and standard deviation found in the tutoring program described above. Given these parameters and a significance level of α=0.05, what is the sample size needed to achieve 80% power?

B. The EGAP calculator gives the total sample size. Assuming half is allocated to treatment and half is allocated to control, how many do you

need in your treatment and control groups? (Round up to the nearest whole number).

c. **Changing significance level.** Statistical tests typically use a 5% significance level but may also opt for a 1% significance level. If you want to be more strict and only allow for a 1% likelihood of obtaining a false positive, what is the sample size needed to achieve 80% power? Does changing your significance level from 5% to 1% (while holding everything else constant) require a larger or smaller sample size? Why?

d. **Changing power level.** Power is typically set at 80%, or 0.8, though in some cases it is increased to 90% to further reduce the risk of a false negative. If you want to have a significance level of 5% but want the rate of true positives (power) to be 90%, what is the sample size needed? Does changing your power from 80% to 90% (while holding everything else constant) require a larger or smaller sample size? Why?

Experimenting With Different Expected Effect Sizes

A new study has come out which evaluates the impact of a similar tutoring program, and the local context is more similar to the context in which you plan to run your study. While the prior study of the tutoring program in Andhra Pradesh led you to believe that a 3.8 point increase in test scores is possible, the more recent study suggests that an impact of half the size (1.9 points) is more reasonable. Continue to work with 80% power and a 5% significance level unless otherwise noted.

e. Intuition: Without going through the calculations, what happens to the minimum sample size needed if the effect is a 1.9 point increase in test scores instead of a 3.8 point increase? Why? How much does the required sample size change (roughly)?

F. Now calculate the minimum sample size that is needed to detect the effect size of 1.9. For now, assume that the more recent study also finds a standard deviation in the underlying test scores of 15.2.
Hint: You might need to change the maximum number of subjects to see the number graphically.

G. Recall that in the first part of the exercise you used an effect size of 3.8 points, based on results from a previous study, to do your power calculations. With the new study suggesting that a 1.9 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?

Variance in the Outcome Measure

After having reviewed the two studies, you conclude that you want to power your study for the smaller of the two effects, that is, an effect of 1.9. However, to increase the power of the study/decrease the required sample size, you decide to collect extensive baseline data in order to minimize the unexplained variance in the outcome measure. Your baseline data indicate that the standard deviation of the test scores in your sample is 7.6, which is half that of the original study.

H. Intuition: Without going through the calculations, what happens to the minimum sample size needed to detect a 1.9 point increase in test scores when the standard deviation of test scores decreases from 15.2 to 7.6? Does it increase, decrease, or remain the same? Why?

I. Having gone through the intuition, now calculate the minimum sample size needed to detect a 1.9 point increase in test scores, given a standard deviation in test scores of 7.6.

Limited Resources

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

with a given sample size (that is, whether we are sufficiently powered to detect a given effect size, conditional on budgetary limitations).

J. You find out that you only have enough funds for a total sample size of 400. Using the estimate of a 1.9 increase in test scores and a standard deviation of test scores of 7.6, 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 400 students? How would you determine this?
Hint: This exercise involves backing out what the power of the test would be if the sample size were 400, the effect size were 1.9, and the standard deviation were 7.6. This can be done by sliding the Power Target slider until the required sample size is 400.

K. If you use the 3.8 point increase in test scores as suggested by the first study, is it worth carrying out the study on just 400 students? What is the power of your experiment? Assume a standard deviation of test scores of 7.6.

L. What are the consequences of running an underpowered evaluation?

Clustered Designs

Thus far, we considered a simple design where we randomized at the individual level, and students were either assigned to the treatment (tutoring) or comparison (no tutoring) condition. However, spillovers could be a major concern with such a design: if treatment and control students are in the same classroom or school, 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 the tutoring program even though they did not participate in it themselves.

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, which is a much 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 the behavior of individuals within a given cluster is called **intra-cluster correlation**, and we need to account for it in our power calculations using the **intra-cluster correlation coefficient (ICC)**. The ICC ranges between 0 (if test scores vary as much within each school as they do across schools) to 1 (if all students within each school have identical test scores and if these are different from test scores in all other schools). Realistically, the ICC will fall somewhere between these two extremes.

High intra-cluster correlation

Low intra-cluster correlation

Let's look at how the required number of subjects changes when using a clustered design. 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 based on a the same set of parameters (the same significance level, treatment effect size, standard deviation of the outcome variable, power target, and sample size), as well as some additional parameters (ICC and number of clusters per arm). Keep the significance

level at Alpha=0.05 and the Power Target at 0.8, the treatment effect at 1.9, and the deviation at 7.6. Keep the number of clusters per arm at 40 for now.

M. 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 the ICC to 1? What happens as you slowly decrease it to 0? Intuitively, why is the green line shifting as you change the ICC?

N. Based on earlier tutoring interventions, your research team has estimated an ICC of 0.11. You need to calculate the total sample size to measure a 1.9 point increase in test scores with a standard deviation of 7.6.
Hint: Check the clustered design box in the EGAP calculator and adjust the intra-cluster correlation bar to 0.11.

O. How many subjects do you need in each of the 80 clusters?

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 to the number of participants in a cluster (e.g., a set number of students in a class). We will now examine how manipulating the number of clusters versus the number of participants per cluster affects power, starting with the number of clusters.

P. Using the same effect size, standard deviation, and ICC as the previous question, 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?

Q. You realize that you misread the earlier studies: it turns out that the estimated ICC is actually 0.07 and not 0.11. Now what would the number of students per cluster and total number of students be if you had 60 schools per arm (or 120 schools in total)?

R. How do your answers here compare to your answers when the ICC was 0.11? Why?

S. Given a choice between offering 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?

Take-up and Attrition (optional - revisit after threats & analysis lecture)

Take-up rates are often imperfect. Being assigned to the treatment group (most often) means that you are offered the program—not that you receive the program with certainty. Not everyone will be interested in taking up the program, and participants who take up the program initially can drop out before its completion. In the case of the tutoring example, imperfect take-up might be caused by a lack of time to participate in the extra tutoring sessions or a fear of the stigma associated with receiving extra tutoring.

Attrition

Attrition occurs when—for any reason—you are unable to collect endline data from participants who were originally included in your sample and randomized into the control or treatment group. Attrition affects the **effective sample size**, i.e., the sample size that determines the power of your study, regardless of how large a sample size you recruited before the start of the study. For example, if you need a sample size of X to detect an effect size of Z with 80% power on a 5% significance level, but your attrition is 25%, you need to recruit a sample of X/(1-0.25). In the case of the tutoring example, attrition might happen because some students move, some might

have dropped out of school, and some who are still enrolled in school might be sick or traveling when the test takes place.

T. The study that found an effect of 1.9 had an almost perfect take-up of the program and no attrition. You expect that your program will have the same effect among the students who take up the program, but you only expect a take-up rate of 75% based on your pilot study. If you assume that everyone who receives the treatment experiences the same benefit in terms of test score increases, what would your expected treatment effect be with a take-up rate of 75%? For simplicity, assume we are again randomizing at the individual level.

U. What sample size would you need in order to have 80% power to detect the effect size you found above if the variance of the outcome is 7.6?

V. Now assume that you expect that your study has an attrition rate of 20%. In order to be able to detect the effect size you found with imperfect take-up, how many people do you need to recruit initially with an expected attrition rate of 20%?

Power Calculations by Hand (Optional - Challenge Problems)

W. **Challenge Question 1**: Redo question **A** by hand, using the formula from the lecture. Continue to assume 80% power and a 5% significance level. (Calculators are allowed!)

$$N = (t_{1-\kappa} + t_{1-\alpha/2})^2 \times \frac{1}{P(1-P)} \times \frac{\sigma^2}{MDE^2}$$

X. **Challenge Question 2**: The EGAP calculator only allows for an even split where 50% of the sample is randomly allocated to the treatment group

and 50% is randomly allocated to the control group. However, studies may want to deviate from a 50/50 split (e.g., perhaps there are enough resources to offer the intervention to more than half the eligible study sample). Use the formula to calculate the sample size needed if you put ⅔ of the group into the treatment, and ⅓ into the control group.

y. **Challenge Question 3**: The power formula looks complicated and often relies on data you might not have yet. See if you can reduce it to a simpler form with just a few assumptions. Most studies have a 50/50 split, and if your outcome is binary, the worst case scenario standard deviation is 0.5. (A binary outcome's standard deviation, $\sqrt{p * (1 - p)}$, is calculated deterministically from its mean, p.) Rewrite the formula with these assumptions so that N is a function of MDE and a constant. Manipulate this further to produce a second formula where MDE is a function of N and a constant.

z. **Challenge Question 4**: Using the formula you just derived, determine the MDE you would be powered to calculate with standard parameters (5% significance level and 80% power) if your outcome of interest is passing or failing a particular class (a binary outcome!), you are randomizing at the individual (student) level, and your sample size is 500 students. What about 1,000 students?

   Note that you can interpret your result as follows: "With a sample size of N students, this study would be powered to detect an impact as small as an X percentage point increase on the likelihood to pass the class."

Tools and Code

Djimeu, Eric W. and Deo-Gracias Houndolo (2016). Power calculation for causal inference in social science: sample size and minimum detectable effect determination. 3ie Working Paper 26. (report and calculator tool).

J-PAL's Sample code on conducting power in Stata and R

Optimal Design program for power calculations and instructions on optimal design.

Resources

Banerjee, Abhijit, et al. Remedying Education: Evidence from Two Randomized Experiments in India (2007). *The Quarterly Journal of Economics*, vol 122 (3): 1235-1264. (evaluation of the Balsakhi tutoring program).

Gelman, Andrew, and Jennifer Hill. (2006). Sample size and power calculations. In *Data Analysis Using Regression and Multilevel/Hierarchical Models*: 437-454. Cambridge: Cambridge University Press.

J-PAL Research Resource: Power calculations

McConnell, Brendon and Marcos Vera-Hernandez (2015). Going beyond simple sample size calculations: a practitioner's guide for power calculations. Institute for Fiscal Studies Working Paper W15/17. (includes sample Stata code)

McKenzie, David (2011). Power Calculations 101: Dealing with Incomplete Take-up.

REUSE AND CITATIONS

> J-PAL. 2023. "Exercise: How to Do Power Calculations." Abdul Latif Jameel Poverty Action Lab. 2023. Cambridge, MA.