# **Power Guide - Alex original, Gareth edits (many based on Winston’s excellent comments)**

# **1 What Power Is**

Power is the ability to distinguish signal from noise.

The signal that we are interested in is the impact of a treatment on some outcome. Does education increase incomes? Do public health campaigns decrease the incidence of disease? Can international monitoring decrease election fraud?

The noise that we are concerned about comes from the complexity of the world. Outcomes vary across people and places for myriad reasons. In statistical terms, you can think of this variation as the standard deviation of the outcome variable. For example, suppose an experiment uses rates of a rare disease as an outcome. The total number of affected people isn’t likely to fluctuate wildly day to day, meaning that the background noise in this environment will be low. When noise is low, experiments can detect even small changes in average outcomes. A treatment effect that decreased the incidence of the disease by 1% would be easily detected, because the baseline rates are so constant.

Now suppose an experiment instead used subjects’ income as an outcome variable. Incomes can vary pretty widely – in some places, it is not uncommon for people to have neighbors that earn two, ten, or one hundred times their daily wages. When noise is high, experiments can struggle to detect effects. A treatment effect that increased workers’ incomes by 1% would be difficult to pick up, because incomes differ by so much in the first place.

A major concern before embarking on an experiment is the danger of a false negative. Suppose the treatment really does have a causal impact on outcomes. It would be a shame to go to all the trouble and expense of randomizing the treatment, collecting data on both treatment and control groups, and analyzing the results, just to have our hypothesis tests be overwhelmed by background noise.

If our experiments are highly-powered, we can be confident that if there truly is a meaningful treatment effect, we’ll see it in our formal tests.

# **2 Why You Need It**

Experimenters guard against false positives with statistical significance tests. After an experiment has been run, we are concerned about incorrectly concluding that there is an effect when---as a matter of fact---there isn’t one.

Power analysis asks the opposite question: supposing there truly is a treatment effect and you were to run your experiment a huge number of times, how often will you get a statistically significant result?

Answering this question isn’t easy. To run a power analysis, the researcher will have to invoke assumptions and rely on a certain amount of informed guesswork. Some of the ingredients that go into making a power calculation are easily knowable---for example, how many subjects can your organization realistically afford to have participate in the study? But others---like what is the smallest treatment effect you’d be willing to see from the experiment?---are trickier. The great benefit of power analysis is that it allows us to see how sensitive the probability of getting significant results is to changes in these assumptions.

Many disciplines have settled on a target power value of 0.80. Researchers will tweak their designs and assumptions until they can be confident that, if there is a treatment effect that is large enough to be of interest, their experiments will return statistically significant results 80% of the time. However, norms are evolving. 80% power means that 20% of the time our experiment will fail to reject a false null hypothesis (a so-called “Type II error”). This represents a pretty big risk, and one that we would almost never tolerate for Type I errors (i.e. erroneously rejecting a true null hypothesis). As a result, some researchers now advocate setting target power at 90% or 95%.

A note of caution: power has to be taken seriously. Null effects from underpowered studies can be hard to interpret. Did the intervention really have no impact? Or is the study just not able to figure it out? Non-null results from an underpowered study can also be misleading: conditional upon being statistically significant, an estimate from an underpowered study probably overestimates treatment effects. Under-powered studies are sometimes based on overly optimistic assumptions about how the study will turn out. A convincing power analysis makes these assumptions explicit. This should protect you from implementing designs that realistically have no chance of answering the questions you want to study.

# **3 The Three Ingredients of Statistical Power**

There are three big categories of things that determine how highly powered your experiment will be. The first two (the strength of the treatment and background noise) are oftentimes given by your experimental environment and can be hard to control. But the last---the experimental design---is almost entirely in your hands. Take advantage of it!

* Strength of the treatment. As the strength of your treatment increases, the power of your experiment increases. This makes sense: if your treatment were giving subjects $1,000,000, there is little doubt that we could discern differences in behavior between the treatment and control groups. In some cases, researchers have the ability to bolster the strength or “dosage” of the treatment. Depending on budget constraints, they might be able to send canvassers to knock on the same doors more than once in a get-out-the-vote experiment. Other times, though, we don’t have as much control as we would like. It is quite common for applied researchers to partner with NGOs or government agencies to evaluate already-existing programs. Here, we may be stuck with evaluating the regimens we are given.
* Background noise. As the background noise of your outcome variables increases, the power of your experiment decreases. We know that some outcomes (e.g. household income) tend to be more variable than others (e.g. people’s height). This is rarely a determining factor for researchers deciding what they should study, nor should it be. But it’s important to be aware of the issue up-front, because there are several things we can do to mitigate the problem of statistical noise. Solutions include gathering covariates about experimental subjects (see below; also *link to covariate adjustment guide*) and trying to keep between-subject heterogeneity to a minimum. Another worthwhile endeavor is to keep your measures of outcomes as clean as possible. One suggestion that has been proposed is to collect and combine multiple measures of the same outcome.[[1]](#footnote-0)
* Experimental Design. Traditional power analysis focuses on one (albeit very important) element of experimental design: the number of subjects in each experimental group. A larger number of subjects increases power. The reason has to do with sampling variation. Suppose we evaluate an intervention to boost students’ test scores. Suppose, too, that we randomly sampled students from a larger population, but were only able to include a small handful of children in the study. In this scenario, there’s a non-trivial chance that we will happen to select a group of relatively high-performing students for the treatment group, and lower-performing students for the control group. If we’d been able to include more students in the sample, this would have been much less likely. For larger samples, we get an increase in power because we can be more confident that any effects we *do* see aren’t just produced by sampling variability. Bear in mind though, that increasing the sample size almost always involves extra costs, in time and money. And despite the emphasis it gets, sample size isn’t everything. There are other elements of the experimental design that can increase power too. How is the randomization conducted? Are there equal numbers of experimental units in treatment and control groups? Will other factors be statistically controlled for? How many treatment groups will there be, and can they be combined in some analyses? Will you conduct a one-sided or a two-sided test?

# **4 Key Formulas for Calculating Power**

Statisticians have derived formulas for calculating the power of many experimental designs. They can be useful as a back of the envelope calculation of how large a sample you’ll need. Be careful, though, because the assumptions behind the formulas can sometimes be obscure, and worse, they can be wrong.

Here is a common formula used to calculate power for a two-sided test[2](https://rawgit.com/egap/methods-guides/master/power/power.html#fn2)

PW=Φ(|μt−μc|N‾‾√2σ−Φ−1(1−α2))β=Φ(|μt−μc|N2σ−Φ−1(1−α2))

* PW is our measure of power. Because it’s the probability of getting a statistically significant result, this will be a number between 0 and 1.
* ΦΦ is the CDF of the normal distribution, and Φ−1Φ−1 is its inverse. Everything else in this formula, we have to plug in:
* μtμt is the expected outcome in the treatment group. Suppose it’s 65.
* μcμc is the expected outcome in the control group. Suppose it’s 60.
* Together, assumptions about μt and μc define our assumption about the size of the treatment effect: 65-60= 5.
* σσ is the standard deviation of outcomes. This is how we make assumptions about how noisy our experiment will be — one of the assumptions we’re making is that sigma is the same for both the treatment and control groups. Suppose σ=20σ=20
* αα is our significance level – the convention in many disciplines is that α should be equal to 0.05. NN is the total number of subjects. This is the only variable that is under the direct control of the researcher. This formula assumes that every subject had a 50/50 chance of being in control. Suppose that N=500N=500.

Working through the formula, we find that under this set of assumptions, β=0.80β=0.80, meaning that we have an 80% chance of recovering a statistically significant result with this design. Click [here for a google spreadsheet](https://docs.google.com/spreadsheets/d/117R4cqKkhX1MFqPIh7Yg2YzjHykxD7WsSLXqhEbD33I/edit#gid=0) that includes this formula. You can copy these formulas directly into Excel. If you’re comfortable in R, here is code that will accomplish the same calculation.

The big unknowns in the formula given above are *sigma*---the standard deviation of the outcome variable---and the acceptable treatment effect size (μt−μc). What should we plug in here?

A few disciplines have settled on standardized benchmarks for what constitutes a “substantial treatment effect,” i.e., one big enough to be of substantive interest. In the field of education policy, for example, a 0.2 standard deviation treatment effect is generally considered meaningful for interventions that try to improve children’s test scores.[[2]](#footnote-1) Where such benchmarks aren’t available, the next best thing to do is to review published studies that investigate interventions similar to yours. If studies generally agree on the substantive impact of an intervention, you probably want to ensure that your own study is sufficiently powered to detect an effect of that size. A final possibility---one appropriate for some but not all applications---is to calculate what size treatment effect would we need to observe for the program to be cost-effective. Whichever line of attack you pursue, be sure to compute power calculations for a range of possible treatment effects.

A slightly different way to think about this issue is to rejig your power analysis and ask: What is the minimum effect size I could detect for a given number of subjects, this type of design, and for this level of significance and power? The answer is known as the minimum detectable effect (MDE) size, and the calculation implies that you have sufficient power to detect a true effect size greater than or equal to the MDE, but you don't have adequate power to detect a true effect size smaller than the MDE. If you're okay with this, then it’s time to move forward with the experiment. If not, modify the design and run the power calculations again.[[3]](#footnote-2)

What about the estimate of sigma? Previous studies---including purely descriptive reports---can be a rich trove of information about the likely standard deviation of your outcome variable. You might also consider running a pilot study on a smaller study sample that can shed light on this question more directly. Again, researchers who are uncertain about sigma may wish to compute power under a range of assumptions about sigma’s value.

# **5 When to Believe Your Power Analysis**

We’ve tried to highlight that all power analyses rely on questionable assumptions. The good news is that it is easy to find out how much your conclusions depend on your assumptions: simply vary those assumptions and see how power changes in response.

This is most easily seen by thinking about how power varies with the number of subjects. A power analysis that looks at power for different study sizes simply plugs in a range of values for N and shows how the power changes.

Using the formula in section 4, you can see how sensitive power is to all of the assumptions: Power will be higher if you assume the treatment effect will be larger, or if you’re willing to accept a higher alpha level, or if your outcome variable has a smaller standard deviation.

# **6 How to Use Simulation to Estimate Power**

Power is a measure of how often, given assumptions, we would obtain statistically significant results, if we were to conduct our experiment thousands of times. The power calculation formula takes assumptions and returns an analytic solution. However, due to advances in modern computing, we don’t have to rely on analytic solutions for power analysis. We can tell our computers to literally run the experiment thousands of times and simply count how frequently our experiment comes up significant.

The code block below shows how to conduct this simulation in R.

The code for this simulation and others is available [here](http://egap.org/content/power-analysis-simulations-r). Simulation is a far more flexible and far more intuitive way to think about power analysis. Even the smallest tweaks to an experimental design may be difficult to capture in a formula (adding a second treatment group, for example) but are relatively straightforward to include in a simulation.

In addition to counting up how often your experiments come up statistically significant, you can directly observe the distribution of p-values you’re likely to get. The graph below shows that under these assumptions, you can get expect to get quite a few p-values in the 0.01 range, but that 80% will be below 0.05.

# **7 How to Change your Design to Improve Your Power**

Judicious tweaks to the design of your experiment can lead to quite dramatic increases in statistical power. As we’ve seen above, an obvious design choice is the number of subjects to include in the experiment. The more subjects, the lower the sampling variability, and the higher the power.

However, the number of subjects is far from the only design feature with upshots for power. Two further issues merit discussion.

* Choice of estimator. Are you using difference-in-means? Will you be doing some transformation, such as a logit or a probit? Will you be using some kind of robust standard error estimator? All of these decisions will make a difference for the power of your experiment. One easy way to think about this is to imagine what command you’ll be running in R or Stata after the experiment has come back; that’s your estimator!
* Randomization Protocol. What kind of randomization will you be employing? Simple randomization gives all subjects an equal probability of being in the treatment group, and then performs a (possibly weighted) coin flip for each. Complete randomization is similar, but it ensures that exactly a certain number will be assigned to treatment. Block randomization is potentially more powerful — it ensures that a certain number within a subgroup will be assigned to treatment. A restricted random assignment rejects some random assignments based on some set of criteria — lack of balance perhaps. These various types of random assignment can often increase the power of an experiment at little or no extra cost.[[4]](#footnote-3)

An alternative to blocking is covariate adjustment: including in the regression model supplementary variables that were measured prior to the rollout of the intervention. If these covariates are strongly related to the outcome, then statistical noise is reduced, and you’ve dramatically increased the power of your experiment. Generally speaking, the covariate with the greatest explanatory power will be the pre-treatment measure of the outcome variable itself, so try to include it where possible. Beware, though, that strongly predictive covariates aren’t always in ready supply. A good place to learn more is Howard Bloom’s paper, “The Core Analytics of Randomized Experiments for Social Research” (pp. 11-14).

Simulation offers a flexible tool for understanding what the precision gains from covariate adjustment and blocking mean for statistical power. Consider the following illustration.

* Suppose we’re studying the effect of an educational intervention on income
* Suppose we have good data on the relationship between two covariates and income: age and gender. In this economy, men earn more than women, and older people earn more than younger people.
* Run a regression of income on age and gender and record the coefficients, using pre-existing survey data (better yet: use baseline data from future participants in your experiment!) \*Generate fake covariate data — N total subjects, but broken up by age and gender in a way that reflects your experimental subject pool.
* Generate fake control data — where the outcome is a function of age and gender according to your regression estimates
* Hypothesize a treatment effect to generate fake treatment data
* Run the experiment 10,000 times, and record how often, using a regression with controls, your experiment turns up significant.

Here’s a graph that compares the power of an experiment that doesn’t control for background attributes to one that does. The R-square of the regression relating income to age and gender is pretty high — around .66 — meaning that the covariates that we have gathered (generated) are highly predictive. For a rough comparison, sigma, the level of background noise that the unadjusted model is dealing with, is around 33. This graph shows that at any N, the covariate-adjusted model has more power — so much so that the unadjusted model would need 1500 subjects to achieve what the covariate-adjusted model can do with 500.

This approach doesn’t rely on a formula to come up with the probability of getting a statistically significant result: it relies on brute force! And because simulation lets you specify every step of the experimental design, you do a far more nuanced power analysis than simply considering the number of subjects.

# **8 Power Analysis for Multiple Treatments**

Many experiments employ multiple treatments which are compared both to each other and to a control group. This added complication changes what we mean when we speak of the “power” of an experiment. In the single treatment group case, power is just the probability of getting a statistically significant result. In the multiple treatment case, it can mean a variety of things: A) the probability of at least one of the treatments turning up significant, B) the probability of all the treatments turning up significant (versus control) or C) the probability that the treatments will be ranked in the hypothesized order.

This question of multiple treatment arms is related to the problem of multiple comparisons. Standard significance testing is based on the premise that you’re conducting a single test for statistical significance, and the p-values derived from these tests reflect the probability under the null of seeing such a large (or larger) treatment effect. If, however, you are conducting multiple tests, this probability is no longer correct. Within a suite of tests, the probability that at least one of the tests will turn up significant even when the true effect is zero is higher, essentially because you have more attempts. Statisticians have proposed a number of possible resolutions to this problem. We discuss these in greater depth in a separate guide on multiple comparisons [*link to multiple comparisons guide*].

# **9 How to Think About Power for Clustered Designs**

When an experiment has to assign whole groups of people to treatment rather than individuals, we say that the experiment is clustered. These designs are very common in educational experiments, where it makes sense to assign classrooms (rather than individual students) to treatment or control. They also feature prominently in research on developing countries, where it is common to assign villages to treatment or control conditions.

As a general rule, clustering decreases your power. If you can avoid clustering your treatments, that is preferable for power. Unless you face concerns related to spillover, logistics, or ethics, take the variation down to the lowest level that you can.

The best case scenario for a cluster-level design is when which cluster a subject is in provides very little information about their outcomes. Suppose subjects were randomly assigned to clusters — the cluster wouldn’t help to predict outcomes at all. If the cluster is not predictive of the outcome, then we haven’t lost power to clustering.

Where clustering really causes trouble is when there is a strong relationship between the cluster and the outcome. For example, in a public health study performed in a developing country, suppose that some villages are, as a whole, much richer than others. Then the clusters might be quite predictive of people’s health outcomes. Clustering can reduce your effective sample size from the total number of individuals to the total number of clusters.

There are formulas that can help you understand the consequences of clustering.[[5]](#footnote-4) While these formulas can be useful, they can also be quite cumbersome to work with. The core insight however is a simple one: you generally get more power from increasing the number of clusters than you do from increasing the number of subjects within clusters. Better to have 100 clusters with 10 subjects in each than 10 clusters with 100 subjects in each. And the statistical complications of analyzing clustered data (e.g., biased estimates of the average treatment effect when clusters vary in size, bias in “robust cluster” estimates of standard errors) become less onerous as the number of clusters increases.

Again, a more flexible approach to power analysis when dealing with clusters is simulation. See the simulations page [*link here*] for starter code.

# **10 Good Power Analysis Makes Preregistration Easy**

When you deal with power you focus on what you cannot easily control (noise) and what you can control (design). If you use the simulation approach to power analysis, you will be forced to imagine how your data will look and how you will handle it when it comes in. You will get a chance to specify all of your hunches and best guesses in advance, so that you can launch your experiments with clear expectations of what they can and cannot show. That’s some work but the good news is that if you really do it you are most of the way to putting together a comprehensive and registerable pre-analysis plan.

An exciting project being pioneered by EGAP members is *DeclareDesign* (*include link*)---a free software package in R that helps researchers to formally characterize their designs according to a set of common standards. The software will take your inputs and then run simulations that can diagnose various properties of the design, including power. Check it out!

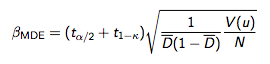
# **Further resources**

An excellent and very readable guide to statistical power is given in Chapter 6 of *Running Randomized Experiments: A Practical Guide* (Princeton, 2013), by Rachel Glennerster and Kudzai Takavarasha.

Bansak (https://arxiv.org/abs/1610.08580) discusses the complications of power analyses in experiments with noncompliance and introduces a new approach.

What software should you use to calculate power? Here (*insert link*) is some detailed advice from the World Bank. If you are a Stata user, the sampsi and sampclus commands are generally sufficient for performing simple power calculations. Optimal Design (include link) is an especially handy tool for running power analysis for more complex group-level interventions. It’s free, but isn’t currently available for Mac users.

MDE FORMULA TO INSERT IN FOOTNOTE ABOVE:



1. For two instructive papers about the usefulness of gathering and combining multiple measures of the same outcome, see Ansolabehere, Rodden, and Snyder, “The Strength of Issues: Using Multiple Measures to Gauge Preference Stability, Ideological Constraint, and Issue Voting” (*APSR*, 2008), and McKenzie, “The Case for More T in Experiments” (*Journal of Development Economics*, 2012). [↑](#footnote-ref-0)
2. However, see Abhijeet Singh’s Development Impact blog post for a thoughtful cautionary note: <http://blogs.worldbank.org/impactevaluations/how-standard-standard-deviation-cautionary-note-using-sds-compare-across-impact-evaluations> [↑](#footnote-ref-1)
3. A popular paper advocating this approach is Howard Bloom, “Minimum Detectable Effects: A Simple Way to Report the Statistical Power of Experimental Designs” (*Education Review*, 1995). A standard formula for calculating the MDE is: INSERT FORMULA HERE [see image at very end of document], where D is the dichotomous treatment variable and V(u) is the variance of the outcome. Note that for an alpha (significance level) of 0.05, t*α/2* is 1.96, while t1−κ is 0.84 for power (κ) set at 0.80. [↑](#footnote-ref-2)
4. To learn more about the mechanics of randomization, see [LINK TO 10 THINGS YOU SHOULD KNOW ABOUT RANDOMIZATION]. See also Chapter 2 of Don Green and Alan Gerber’s book, *Field Experiments: Design, Analysis, and Interpretation* (W.W. Norton, 2012) [↑](#footnote-ref-3)
5. For an extended discussion, see Andrew Gelman and Jennifer Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Model* (Cambridge, 2006), pp. 447-449. [↑](#footnote-ref-4)