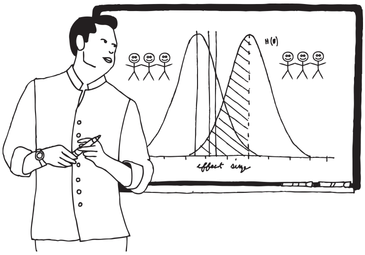
6 Statistical Power



In this chapter we explain what power is, how it is determined, and how we can use power analysis to determine what sample size we need, the level at which to randomize, how many different treatment groups we can test, and many other design issues. It contains the following modules:

MODULE 6.1 The Statistical Background to Power

This module covers the basic statistical concepts we will use in our discussion of power, including sample variation, standard errors, confidence intervals, and hypothesis testing.

We have completed our evaluation of a program designed to improve test scores in a poor community in India. We find that the average score on the test we administer is 44 percent in the treatment group and 38 percent in the comparison group. Can we be reasonably confident that this difference of 6 percentage points is due to the program, or could it be due to chance? We perform a statistical test designed to check whether the difference we observe could be due to chance. We find that the difference between the treatment and comparison groups is not statistically significant, that is, there is a reasonable likelihood that the difference is due to chance. The program was inexpensive, so an improvement of 6 percentage points in test scores would be considered a great success and possibly a target for scale-up and replication, if we could be reasonably confident the difference was due to the program. But we cannot rule out that the difference is due to chance because our evaluation produced a result that was not very precise. A wide range of program impacts are plausible from our results: we can be confident only that the impact of the program was somewhere between –2 and +14 percentage points on the test we administered. Our evaluation was not designed well enough to avoid such a wide range of plausible values. It did not have enough statistical power.

Statistical power can be thought of as a measure of the sensitivity of an experiment (we provide a formal definition at the end of this module). Understanding power and doing power analysis is useful because it helps us decide whether to randomize at the individual or the group level, how many units (schools, communities, or clinics) to randomize, how many individuals to survey, how many times to survey each individual over the course of the evaluation, how many different program alternatives to test, and what indicators to use to measure our outcomes. It helps us design more sensitive evaluations.

Before we formally define power and discuss what determines it, we need to go over some of the basic statistical concepts that we will need to determine the power of an evaluation, in particular sampling variation, standard errors, critical values, and hypothesis testing.[1]

Sampling variation

When we want to measure characteristics of a population, we usually do not collect data on everyone in the population. We can get a pretty accurate picture of what the population looks like by taking a random sample and collecting data on that sample. For example, if we want to know the average test score of a third grader in Mumbai, we do not have to test every child in the city. We can take a random sample of children and assume that they will be representative of the total population of third graders in Mumbai.

[1] A more thorough but still accessible introduction to the statistical concepts discussed here is given by William Greene in *Econometric Analysis* (New Jersey: Prentice-Hall, 2011).

But how many children should we test? And how accurate will our estimate of learning be? Imagine we choose a random sample of 200 children from across Mumbai to test. There is some chance that we will include some very high-scoring children and very few low-scoring children. We will then overestimate the true level of learning. There is also a chance that we will pick particularly poorly performing children and underestimate the true level of learning. Because we are selecting our sample at random, the chance of underestimating should be the same as the chance of overestimating the true level of learning, but that does not mean that we are guaranteed to get exactly the true learning level if we take a sample of 200 children. We will get an *estimate* of the true mean level of learning in Mumbai (the population mean) by taking the mean of learning in our sample (the sample mean). If we take a sample multiple times, we will get a slightly different estimate each time. If we do this and plot the frequency of the number of times we get different estimates, we will end up with something like Figure 6.1.

The sampling variation is the variation between these different estimates that we get because we are testing only a small proportion of the total population. There are other possible sources of variation between our estimate of the mean and the true mean in the population. For example, our test might not be administered correctly and give us a bad measure of learning. But in this chapter and when we perform power analysis, we are concerned only about sampling variation.

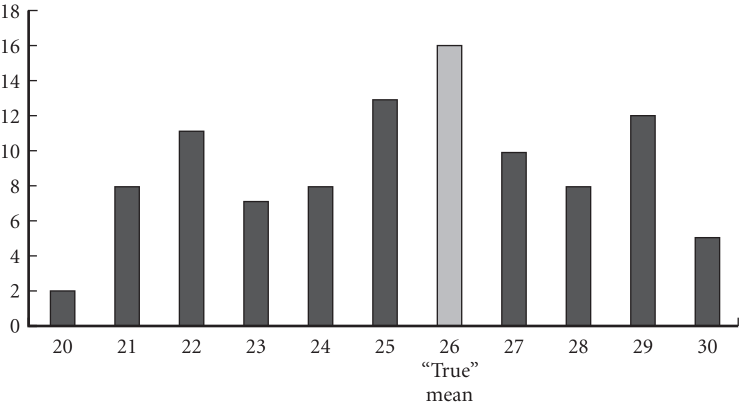


FIGURE 6.1 Frequency of 100 estimates of average learning levels in Mumbai

The extent to which our estimates will vary depending on whom we draw in a particular sample, and thus how confident we can be that our estimated mean is close to the true mean, depends on the variation in test scores in the population as a whole and on how many people we sample. The more people we sample, the more likely that our estimated (sample) average will be close to the true (population) average. In the extreme, if we sample everyone in the population, the sample mean and the population mean will be identical and there will be no sampling variation. In the example below we assume that the sample is small relative to the full population.

The more dispersed the scores are in the population, the more likely that we will include in our sample a child who is far from the mean. Therefore, the less dispersed the sample, the fewer people we have to sample to be confident our estimate is close to the true mean. In the extreme, if everyone in the sample were identical (there was zero variance in the population), we would have to sample only one person to know the true mean.

Standard deviation

A useful measure of the extent of dispersion in the underlying population is the *standard deviation, sd.* To calculate a standard deviation of test scores, we take every third grader in Mumbai, calculate the mean test score for the whole sample, and then for each child calculate the difference between that child's test score and the sample mean and square this difference. Then we add up all the squared differences and divide by the number of children in Mumbai. Finally we take the square root. The algebraic formula for a standard deviation is given as

$$sd = \sqrt{\frac{\Sigma_{i=1}^{n} (x_i - \bar{x})^2}{n}}.$$

Standard error

Remember that if we take a sample of 200 children and calculate our estimate many times we will get a distribution of estimated means. We have said that the dispersion of our estimate of the true mean will depend on two things—the size of the sample of children we test and their dispersion in the test scores of the underlying population. We need to make this relationship a bit more formal. The *standard error* (*se*) measures the standard deviation of our estimate. But we don't usually take many estimates and so can't calculate the standard deviation of the estimates as we did when we calculated the standard deviation of the test scores of children. But if we know the dispersion of the test scores of children themselves, we can use this to calculate the likely dispersion of the estimates. Specifically, the standard error of the estimate is equal to the standard deviation of the population divided by the square root of the sample size:[2–4]

$$se = sd/\sqrt{n}.$$

Central limit theorem

Under certain assumptions (which hold in most of the situations we deal with in this book), the central limit theorem applies.[5] This theorem states that the distribution of the estimated means is close to normal as long as our estimate is based on a large enough sample. A normal distribution is a bell-shaped curve with a particular shape in which approximately 68 percent of all estimates will be within 1 SD of the mean. This is true even if the underlying distribution is not normal. This is important because many of the variables we will use will not be normally distributed. For example, because test scores usually cannot go below zero, we may have a cluster of scores and zero, and the distribution of scores may not be symmetrical around the mean. That our estimate is normally distributed is very useful because if we know the standard deviation of a normal distribution, we know the percentage of times an estimate will fall any distance away from the mean. For example, in a normal distribution we know that approximately 68 percent of all the estimates will be within one standard deviation of the true value and that approximately 95 percent of all estimates will be within two standard deviations of the true value. The standard deviation of the distribution of estimates of the mean is the

[2] Technically this holds true when we are sampling a small proportion of the population. Remember that if we sample the whole population we don't have any sampling error and the standard error is zero.

[3] For a detailed algebraic derivation of the standard error of an estimate, see John A. Rice, *Mathematical Statistics and Data Analysis* (Belmont, CA: Wadsworth and Brooks, 1988), 182–184.

[4] We don't actually observe the standard deviation of the population, so we usually estimate this using the standard deviation of the sample.

[5] A key assumption for the central limit theorem to hold is that the population has a mean and a variance, which is true if the population is finite. A formal exposition of the central limit theorem and a discussion of its practical applications can be found in Rice, *Mathematical Statistics and Data Analysis,* 161–163.

standard error; this means that there is a 95 percent chance that our estimate of the population mean will be within two standard deviations of the true mean.[6]

Confidence interval

When we calculate an estimated mean we get a single number, sometimes called a single point estimate. But we know that there is some uncertainty around this estimate. A *confidence interval* is a range of estimates around a single point estimate. When poll results are reported in the run-up to an election, they are reported with a margin of error, reflecting this uncertainty. For example, we might hear that support for a given candidate is 45 percent ± 3 percent. In other words, the confidence interval is between 42 and 48 percent.

Confidence level

A confidence level will often be reported alongside the confidence interval and the point estimate. A 95 percent confidence level means that if we constructed 100 confidence intervals we would expect 95 of them to contain our parameter.

Hypothesis testing

Measuring differences in averages

Until now we have been drawing just one sample from a population and asking whether our estimate (the average of that sample) is close to the true average of the population. But when we perform a randomized evaluation we are drawing two samples and comparing them to each other. Instead of being interested in estimating the mean of one population, we are interested in estimating the difference between a large hypothetical population that is treated and a population that is untreated.[7] We can estimate this by comparing a randomly chosen sample that is treated and a randomly chosen sample that is not.

[6] Unfortunately we never observe the true SD (unless we measure the entire population). We therefore use the SD in our sample as an estimate of the SD in our population. However, this adds another layer of uncertainty to our estimate that we need to account for. Using the *t*-distribution allows us to account for this additional uncertainty. The *t*-distribution is similar to the normal distribution except that it has fatter tails (in other words, it has a larger number of more extreme values) than the normal distribution and it follows a different distribution depending on the size of the sample. With large sample sizes the *t*-distribution approximates the normal distribution.

[7] We say hypothetical here because our statistical tests assume that we are testing two small random samples from two large populations, one that received the program

To tell whether the program affects test scores, we want to see whether the mean of the treatment sample is different from the mean of the comparison sample; in other words, our estimate of interest is now the difference between the two samples. Fortunately, just as the average of a random variable has a distribution that approaches normal, so does the difference between two averages. In other words the distribution of our estimate will asymptotically approach a normal distribution—that is, it will approach a normal distribution as we run our experiment thousands of times. We can still calculate the standard error of the estimate in the same way and our standard error can give us a confidence interval around our estimate.

The null hypothesis and the research hypothesis

When we estimate the impact of our education program in Mumbai we start with a hypothesis. Our hypothesis is that the program increases test scores and therefore that the average test score of the group exposed to the program is higher than that of the group not exposed to the program:

Hypothesis: Remedial education led to higher test scores.

By convention, every study proposes two hypotheses, the null hypothesis and the research (or alternative) hypothesis. The *null hypothesis* states that the treatment effect is zero, in other words, that the test scores of those exposed and not exposed to the program are the same. The null hypothesis is conventionally denoted H_0:

H_0: Treatment effect = 0.

We sample a subset of the population and expose it to the program while not exposing another subsample. Given that these samples are chosen at random from the same population, without the program these two groups would, on average, have the same test scores. The null hypothesis states that we do not expect any difference between the treatment and comparison groups. Of course, even if the null

and one that didn't. Sometime this is the case. For example, when we randomize at the group level we often test only a small proportion of those who receive the program. In individual-level randomization, however, we often test everyone who received the program. In this case we think of our treatment group as a sample of what we would have seen if we had treated a much larger group of individuals.

hypothesis is true, we can find that in a particular sample the mean test score is 44 percent and in another sample it is 38 percent. The null hypothesis is not inconsistent with an observed difference of (say) 6 percentage points between the two samples; it says that this difference is due to chance (as a result of sampling variability) and that if the program were scaled up it would not affect the average test scores of the population of Mumbai students.

The *research hypothesis* says that the treatment effect is not zero. It is also called the alternative hypothesis and is designated H_1:

H_1: Treatment effect $\neq 0$.

This states that the means of the underlying populations from which the two samples were drawn are not the same. This is the most common form, but the research hypothesis can be more elaborate and say something about the direction of the underlying difference—that one mean is smaller or larger than the other:

H_1: Treatment effect > 0, or else H_1: Treatment effect < 0.

Irrespective of the form it takes, the research hypothesis is the logical opposite of the null hypothesis. (Note that if our research hypothesis is that the program has a positive effect, the null hypothesis is the logical opposite, that is, that the difference is zero or less than zero. Traditionally, we stick to a null hypothesis of zero effect and a research hypothesis of an effect that could be positive or negative, because we want to be able to test whether the program does harm.)

The practice in hypothesis testing is not to test our main hypothesis directly but instead to test its logical opposite. Because the two hypotheses are logical opposites, they cover all the possible inferences that we can reach in a statistical test. So we *test the null hypothesis.* A key notion related to this type of test is that of the "*p*-value," which is the probability of obtaining outcomes such as those produced by the experiment had the null hypothesis been true. A *p*-value of below 0.05 implies that there is only a 5 percent chance or less that an outcome was generated under the null hypothesis. This suggests that we should reject the null hypothesis. If we "reject" the null hypothesis of zero effect, we conclude that the program had an effect. We say that the result is "significantly different from zero." If we "fail to reject" the null hypothesis, we need to ask ourselves. "Is it really true that there is zero effect, or is it simply that our experiment does not have enough statistical power to detect the true effect?"

Statistical inference and statistical power

The program is hypothesized to change the outcome variable (test scores). There is an underlying truth, that the program either would or would not change test scores in the population of students in Mumbai. But we never observe the underlying truth; we observe only the data from our samples. We test whether the means of treatment and comparison samples are statistically different from each other. (In Chapter 8 we go over in detail how we test whether the treatment and comparison samples are different from each other.) But we must remember that statistical tests are not infallible. They only ever give us probabilities because of sampling variation.

The four possible cases

Both the statistical test and the underlying reality have an either/or quality. For the statistical test, either the difference is significant (we reject the null hypothesis) or the difference is not significant (we fail to reject the null hypothesis). For the underlying reality, either there is a treatment effect or there is no treatment effect. This makes four cases possible (Figure 6.2).

1. False positive We have a *false positive* when we find a statistically significant effect even though the program did not actually have a treatment effect. We wrongly infer that the program had an effect. (We reject the null hypothesis when in fact the null hypothesis is true.)

If the program has no effect, the probability of making a false positive error—also called a Type I or alpha error—is called the significance level and denoted α. The convention is that it should be no larger than 5 percent. This says that if the treatment effect is zero—if the null hypothesis is true—the probability that the observed difference is due to chance alone is 5 percent or less. In other words, we have a 5 percent rate of false positives.

2. True zero We have a *true zero* when we find no statistically significant effect and there truly is no treatment effect. We correctly fail to reject the null hypothesis and infer that the program does not work.

The probability that we do not find a statistically significant effect if the treatment effect is zero is called the confidence level. It is the com-

	The underlying truth: Is there a treatment effect?	
Statistical test: Is the observed difference statistically significant?	Treatment effect (H_0 false)	No treatment effect (H_0 true)
Significant (Reject H_0)	True positive Probability = $1 - \kappa$	False positive Probability = α Type 1 error
Not significant (Fail to reject H_0)	False zero Probability = κ Type II error	True zero Probability = $(1 - \alpha)$

In reality, one thing is true. Either there was a treatment effect or there was no treatment effect. But when we combine the possibility of two underlying realities with two possible outcomes from our statistical tests, there are four possible cases we could face with given probabilities: (1) false positive with probability α; (2) true zero with probability $(1 - \alpha)$; (3) false zero with probability κ; and (4) true positive with probability $(1 - \kappa)$. Power is the probability of attaining a true positive. It is the probability that we will detect a treatment effect when there is one.

FIGURE 6.2 Four possible cases, only one of which is true

plement of the significance level. In other words, it is equal to (1 – α). Testing at the 5 percent level, then, gives a confidence level of 95 percent, which says that if the true effect is zero there is a 95 percent probability that we will fail to reject the null hypothesis.

3. False zero We have a *false zero* when we do not find a significant treatment effect (fail to reject the null hypothesis) even though there truly is a treatment effect. We wrongly infer that the program does not affect test scores.

The probability of making a false zero error—also called a Type II error—is denoted here as κ. Some people use a rule of thumb that studies should be designed to have a κ of 20 percent or less. κ can be defined only in relation to a given treatment effect. In other words, a κ of 20 percent means that if the true treatment effect is of a given level, there is a 20 percent chance of failing to find a significant effect. The level of the treatment effect that we want to design our study to be able to distinguish from zero is a whole separate topic that we cover in great detail in the next module.

Power is the probability that we will not make a Type II error. In other words, it is the probability that (if the true effect is of a given size) we will find an effect that is statistically different from zero. Power is therefore shown as (1 – κ). If we aim for a κ of 20 percent, we are aiming for 80 percent power.

4. True positive We have a *true positive* when we find a statistically significant effect (reject the null hypothesis) and there truly is a treatment effect. We correctly infer that the program works.

The probability of finding a true positive when there really is an effect of a given size is (1 – κ). This probability is our statistical power. If the probability of detecting a significant effect (when there is one) is 80 percent, we say that we have 80 percent power.

What level of certainty do we need?

By convention, the significance level—the chance of finding a false positive—is set at just 5 percent.[8] At this low probability, if there is no treatment effect, we are highly unlikely to conclude that there is one.

[8] See, for example, Howard S. Bloom, *Learning More from Social Experiments* (New York: Russell Sage Foundation, 2005), 129, as well as the discussion in Alan Agresti and Barbara Finlay, *Statistical Methods for the Social Sciences* (New York: Prentice-Hall, 1997), 173–180.

Usually the chance of a false negative (failing to reject the null hypothesis when the truth is that the program had an effect of a given size) is set at 20 percent (or sometimes 10 percent). Typically we worry more about false positives than about false negatives. If we set the chance of a false negative (for a given effect size) at 20 percent, we are saying that we want 80 percent power, that is, we want an 80 percent chance that (for a given effect size) we will find a statistically significant effect.

Module 6.1 summary

Statistical background to power

- Sampling variation is the variation in our estimate due to the fact that we do not sample everyone in the population: sampling variation is the reason we do power calculations.
- Standard deviation is a measure of the dispersion of the underlying population.
- Standard error is the standard deviation of our estimate and is related to the size of our sample and the standard deviation of the underlying population.
- A confidence interval gives a range of values around a given estimate.
- Statistical significance uses standard errors to say whether we can be confident that the true estimate is different from a given value: for example, whether we can be confident that the estimate is not zero.

Hypothesis testing

- The null hypothesis is the opposite of our research hypothesis: if our research hypothesis is that we expect the treatment group to have different outcomes from the comparison group, the null hypothesis says that there is no difference between the two groups.
- Traditionally we test the null hypothesis, not the research hypothesis.
- Type I error is rejection of the null hypothesis when in fact it is true (false positive).
- The significance level (traditionally 5 percent) is the chance we are willing to take that Type I errors will occur.

- Type II error is failure to reject the null hypothesis when in fact there is a difference between the treatment and comparison groups (false negative).
- Power measures the probability that we will avoid Type II errors, and we conventionally design experiments with 80 percent or 90 percent power.

MODULE 6.2 The Determinants of Power Explained Graphically

This module explains which factors determine the power of an experiment and how different features of the evaluation design and analysis can affect power. In this module we show graphically and intuitively how power is related to its key components. In the next module we provide a more formal, algebraic explanation of power determinants.

At the end of the previous module we defined power as the probability that (if the true effect of a program is of a given size) we will find a statistically significant effect. To understand the determinants of power we need to return to the issue of sampling variation (discussed in the previous module), in particular the question of how we can tell whether a difference between our treatment and comparison groups is due to chance. In other words, could the difference we see between treatment and comparison groups be due to the fact that we happened to sample individuals with slightly different characteristics in our treatment and comparison groups?

Figure 6.3 is a frequency chart with the results from one randomized evaluation of an education program in India. It shows the number of children who achieved different test scores in both the treatment and the comparison groups on the same chart. It also shows the estimated means for the treatment and comparison groups.

We can see that the mean of the treatment group is slightly higher than the mean of the control group. But is the difference statistically significant? To find this out we must examine how accurate our estimate of the difference is. Accuracy in this case means how likely our estimate is to be close to the true difference.

If we run a perfectly designed and executed experiment an infinite number of times, we will find the *true difference.* Each time we run the experiment, we will get a slightly different estimate of the difference

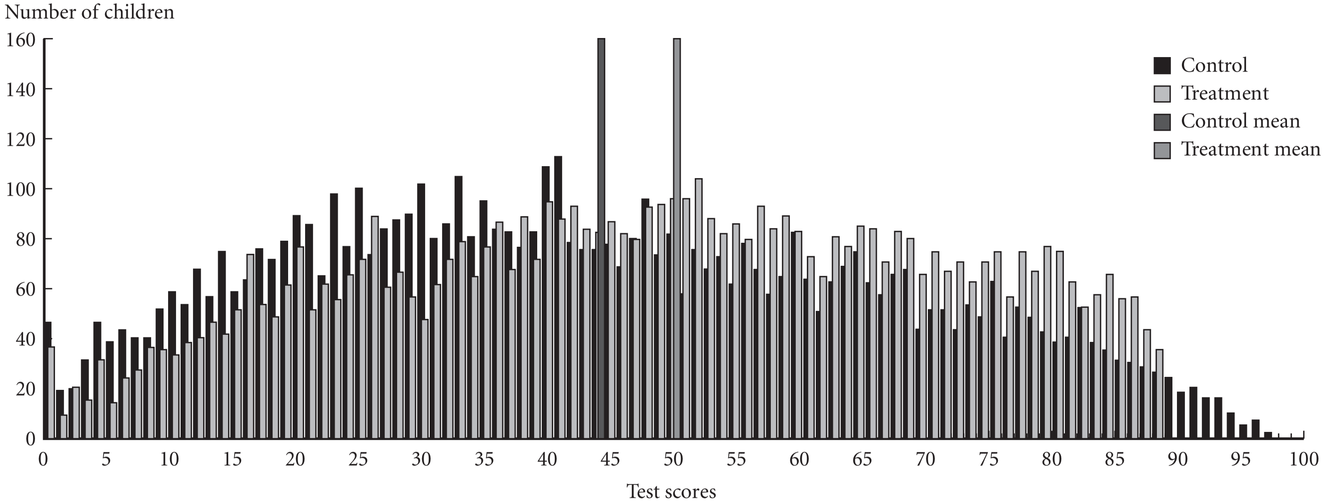


FIGURE 6.3 Test scores of treatment and comparison groups in an educational program

between the treatment and comparison groups and thus of the impact of the program. One time we run the evaluation we might by chance get most of the really motivated children in the treatment group and thus overestimate the impact of the program because these children would have improved their test scores a lot even without the program. Another time we run it, the motivated children might be mostly in the comparison group and we will underestimate the true impact of the program. Because we are running a randomized evaluation and therefore have no selection bias, the average of all our estimated program impacts, if we run the evaluation many times, will be the same as the true impact. But we cannot guarantee that every individual evaluation will give us exactly the same impact as the true impact. Sometimes, by chance, we will get an estimated difference that is much higher or lower than the true difference.

If the true difference (the true effect size) is zero, a very large number of experiments will generate a bell-shaped curve of estimated effect sizes centered on zero (H_0). In other words, the most common result of the many experiments we run will be a difference near zero. But the tails of the bell curve show the number of times we will get an estimated effect that is far from zero.

It is useful to introduce a little bit of notation here. Let us use β to represent the true effect size. If we hypothesize that the true effect is zero, we say that we assume that $\beta = 0$. What if the true effect size is β^*, that is, what if $\beta = \beta^*$? Then running a very large number of experiments will result in a bell-shaped curve of estimated effect sizes centered on β^* (Figure 6.4).

In this example there is considerable overlap between the curves for $\beta = 0$ and $\beta = \beta^*$. This means that it is going to be hard to distinguish between these two hypotheses. Imagine that we run an evaluation and get an estimated effect size of $\hat{\beta}$. This effect size is consistent both with the true effect size's being β^* and with the true effect size's being zero. The heights of points A and A′ show the frequency with which we would get an estimated effect size of $\hat{\beta}$ if the true effect size was β^* or zero, respectively. A is higher than A′, so it is more likely that β^* is greater than zero, but there is still a reasonable probability that the true effect size is zero. The more the curves overlap, the more likely we are to get a result that could come from either hypothesis and thus the less likely we are to be able to tell whether there was an effect. If there are lots of estimated effects that have a reasonable prob-

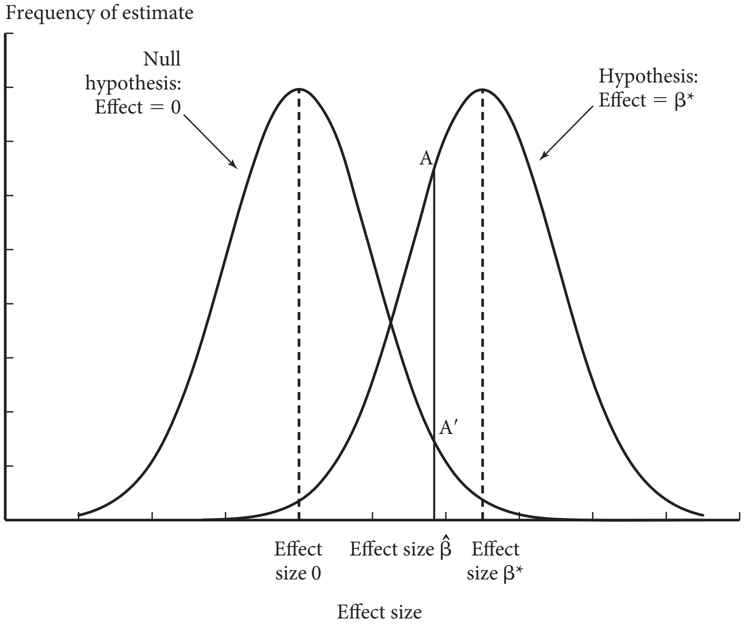


FIGURE 6.4 Frequency of estimated effect sizes from running multiple experiments under both the hypothesis $\beta = 0$ and the hypothesis $\beta = \beta^*$

ability of being generated by the null hypothesis or by the $\beta = \beta^*$ hypothesis, we have low power.

We will always have to live with some overlap between our two bell curves because there is always a small chance of an estimated difference that is far from the true difference. How much overlap are we willing to live with? As discussed in Module 6.1, traditionally we are willing to reject the null hypothesis if there is only a 5 percent chance that it is true. This occurs if the estimated treatment effect falls in the thin tails of the distribution of likely effect sizes under the null hypothesis, above (or below) the 95 percent significance level. The value of the estimated treatment effect that exactly corresponds to the 95 percent significance level is called the *critical value*—anything above this is statistically significantly different from zero, while anything below it is not. This 95 percent critical value is shown in Figure 6.5. We can see that, in the example drawn here, the bulk of the mass of the research hypothesis curve (on the right) is above the critical value of 95 percent for the null hypothesis curve. This means that we are likely

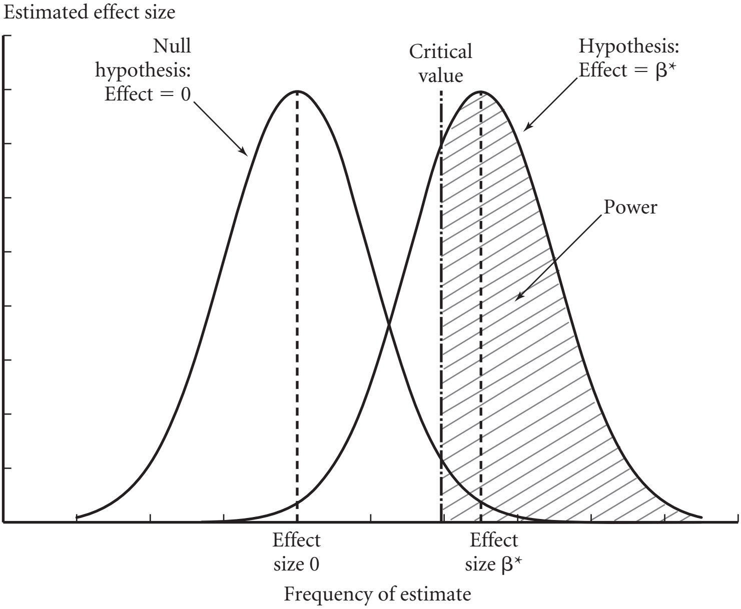


FIGURE 6.5 Distribution of estimated effects with critical value

to find a significant effect if there truly is one of the anticipated size. What is the precise chance that we will be able to reject the null hypothesis if the true effect is β*? This is given by the percentage of the research hypothesis curve that is above the critical value for the null hypothesis. In Figure 6.5, this is the percentage of the total area under the research hypothesis curve that is shaded. This percentage is the definition of the power of the experiment. In Figure 6.5, 70.5 percent of the area under the curve is shaded, meaning that there is 70.5 percent power.

Figure 6.5 shows a *one-sided test;* in other words, it tests whether the program has a positive effect. This test is based on our assuming that a negative effect is not possible. Normally we do not want to impose this assumption, so we test the null hypothesis of no effect against the research hypothesis of either a positive or a negative effect. Although there is usually no strong reason that the positive minimum detectable effect (MDE) size should be the same as the negative MDE, we test symmetrically. In other words, the symmetry of the normal distribution implies that if we want to be able to pick up a positive

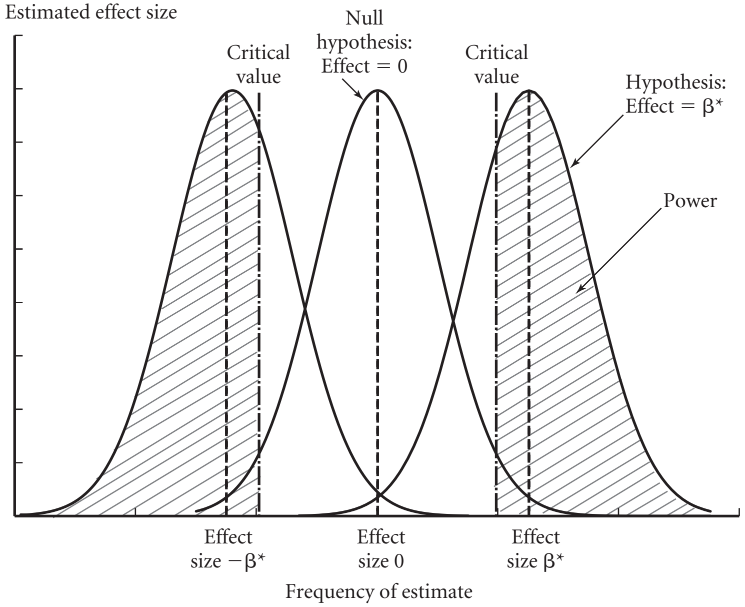


FIGURE 6.6 Distribution of estimated effects with critical value for a two-sided test

effect of magnitude β (with 80 percent power) we should use a two-sided test, which also has 80 percent power to pick up an effect size of –β with 80 percent power. Note, however, that the MDE is always just a point on a continuum; in other words, if we do this we will have some power to pick up a less negative effect than –β. We will have more than 80 percent power to pick up an effect that is more negative than –β. This is shown in Figure 6.6. For the rest of this section, however, we show only a positive research hypothesis in our graphs because it makes the figures easier to follow.

Minimum detectable effect size and power

In Figure 6.7 we can see that power is greater and the overlap between the curves is smaller the farther the two curves are away from each other (i.e., the greater the true effect size). In the case shown here, the true effect size is twice what it is in Figure 6.4 and power is 99.9 percent.

So power is related to the true effect size, but we never know what the true effect size is. We are running the experiment to estimate it. So

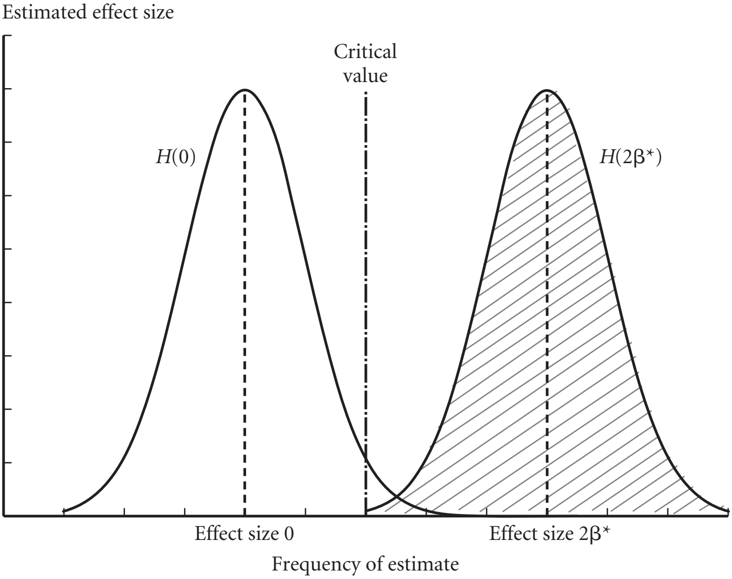


FIGURE 6.7 Distribution of estimated effects with a doubled effect size

how do we calculate power? The answer is that we draw up hypotheses. We say, "What would be the power if the true effect was β^* or $2\beta^*$?" What a power calculation can answer is "What chance do we have of finding a statistically significant effect if the true effect size is β^*?" Or "Given this sample size, what is the minimum level of true effect at which we would have enough power to reject the null hypothesis using a 95 percent significance cutoff?" This hypothesized β is our MDE (see Module 6.1). Choosing the right MDE is a critical part of power analysis, and in the next module we discuss how to choose one in more detail.

Residual variance and power

The hypothesized effect size determines how far apart the two bell curves are. But what influences the shapes of our bell curves of estimated effect size? The width of the bell curves reflects the variance in our estimate of the difference. As we discussed in Module 6.1, variance in our estimate comes from sampling variance; in other words, it comes from the fact that we are testing only a sample of the population, and we may by chance choose people with different outcomes for our treatment group and our comparison group. The greater the variance in the underlying population, the greater our sampling variance will be (for a given sample size). In the extreme case, if all the children in the population had identical test scores, every time we drew a sample of them we would get the true value of test scores in our population, and we would know that any difference we saw between the treatment and comparison groups was due to the program and not to sampling variation. The larger the sample size, the lower the variance in our estimate (i.e., the narrower our bell curves).

Some of the differences in test scores between children will be correlated with observable factors such as age or parental education. If we include age and parental education as "control variables" in our final analysis, we sharply reduce the likelihood that we will misestimate the true effect if, by chance, we have more children with highly educated parents in our comparison group than in our treatment group. The most important control variable to include is usually the baseline level of our final outcome variable. Our estimate of the effect size is more precise, our bell curve is narrower, and our power is higher. Thus our power depends on the residual variance after we have controlled for other available variables. Residual variance depends on the underlying variance of the population and the extent to which this variance can be explained by observable factors for which we have data that we intend to use as controls in the final analysis. In the equations that follow (Module 6.3) we denote residual variance as σ^2.

In the case study at the end of this chapter we show just how important control variables can be for power. For example, including baseline test scores in an education program designed to improve test scores can reduce residual variance by up to 50 percent, allowing for a sharp reduction in sample size for a given power.

When we are performing a group-level randomization, the variance of our estimate of the effect will depend on a few other factors. We discuss these when we discuss power for group-level randomization below.

Sample size and power

The sample size of the experiment will also affect the width of the bell curves. The larger the sample, the more accurate the estimated difference will be and thus the narrower the bell curves. Figure 6.8 shows how in the remedial education example above, the shape of the bell curve of estimated effect size changes as the sample size increases. The

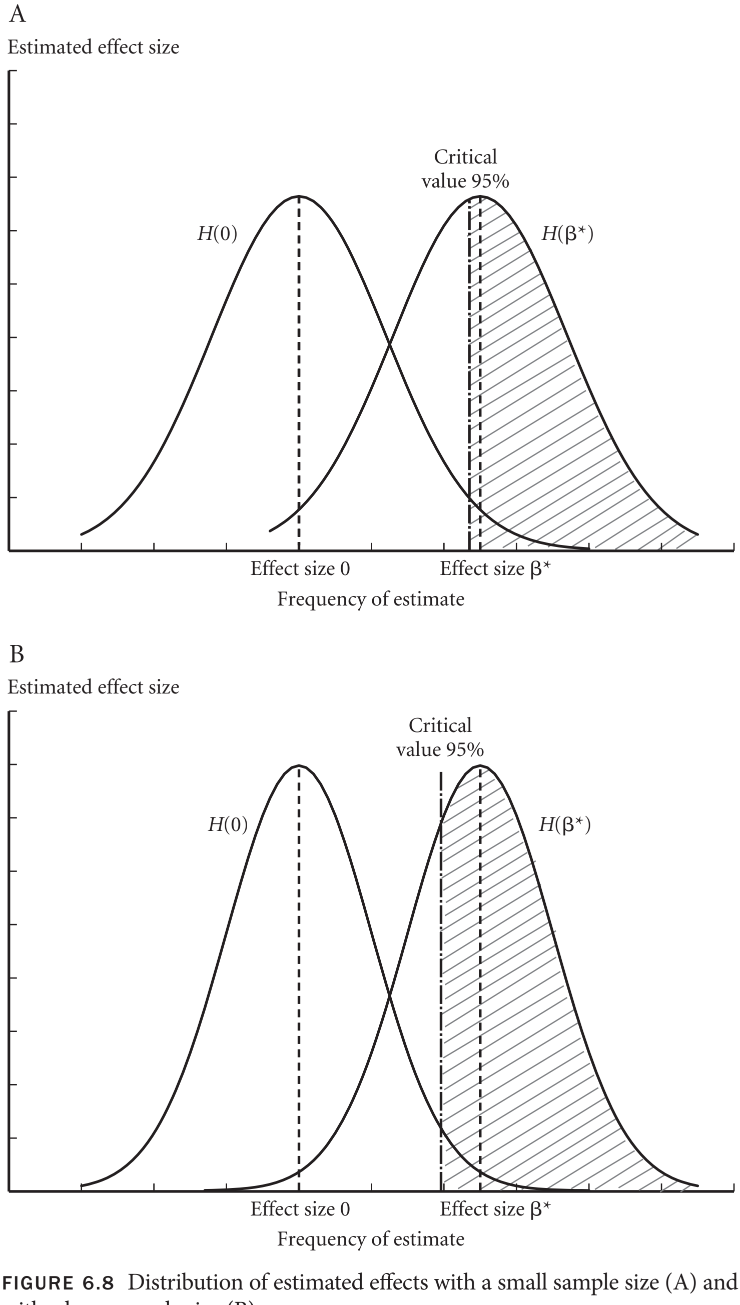


FIGURE 6.8 Distribution of estimated effects with a small sample size (A) and with a large sample size (B)

larger the sample, the more accurate our estimate, which we can see in the narrower bell curve and higher power. We denote sample size as *N*.

Significance level and power

Power also depends on what we decide should be the critical value for significance. For example, it is traditional to use a 5 percent cutoff for statistical significance. A 5 percent significance level means that we reject the null hypothesis only if there is a 5 percent chance or less that the result is due to chance. There is nothing magical about 5 percent, and some academic papers show results even though they are significant at only the 10 percent level (showing a 10 percent probability that the result is due to chance). Loosening the significance level moves the threshold to the left (as shown in Figure 6.9). This increases the shaded area to the right of the critical value, increases the chance that we will reject the null hypothesis (i.e., we will find a statistically significant difference if there is one), and increases the level of power. However, we should realize that if we loosen significance to increase power (and therefore reduce the likelihood of Type II false zeros) it comes at the cost of increasing the chance of false positives. In the power equation, the critical value is denoted as α and the proportion of the curve to the left of the critical value is given by t_{α}. We show how the critical value changes with α in Figure 6.9.

It may seem counterintuitive that moving from a 5 percent to a 10 percent cutoff level for significance increases power, so let us use an analogy. Imagine that we are the jury in a murder trial and we have to decide whether Xi is guilty or not guilty of poisoning Yi based on evidence presented in court. The prosecution submits 20 pieces of evidence. Because not one of us on the jury was present when the poisoning took place, there is always a chance that we could reach the wrong verdict, either saying that Xi is not guilty when he really did poison Yi or saying that he is guilty when he really is innocent. Recognizing this, we decide to set a standard for how stringent we should be. We say that we will convict Xi only when the probability that we could be wrong is 5 percent or less; if it's any higher we will say that he is not guilty. That is, each of us will vote to convict only if we find at least 19 of the 20 pieces of evidence put forward convincing and at most 1 of them unconvincing. If 2 are unconvincing, we vote not guilty. At the 10 percent level, we will need at least 18 convincing pieces of

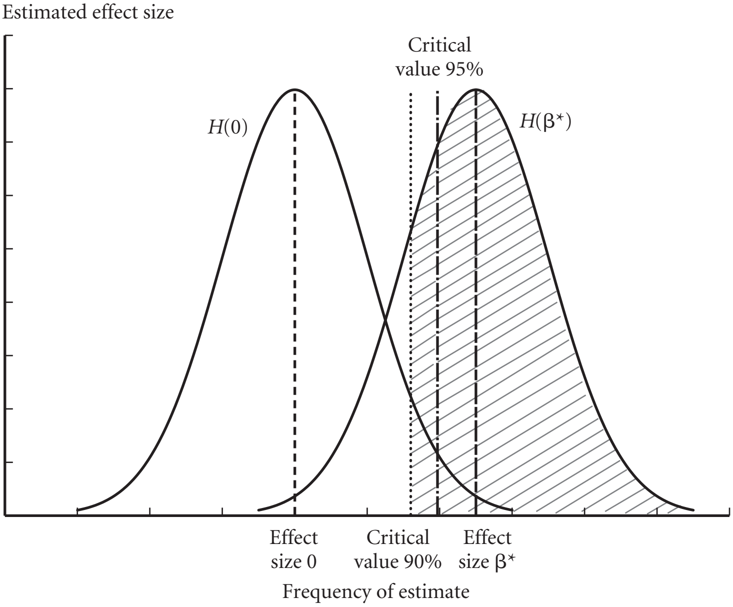


FIGURE 6.9 Distribution of estimated effects with two thresholds for the critical value

evidence and allow for two unconvincing pieces. We will vote not guilty if 3 are unconvincing. Power is the probability that we will find Xi guilty when he really did commit the murder. We are more likely to find him guilty the more pieces of unconvincing evidence we choose to ignore. At the 5 percent level, we will find him guilty even if one piece of evidence is unconvincing, but at the 10 percent level we will do so if 2 pieces are, at the 30 percent level if 6 pieces are, and so on. The less stringent we are, the more likely we are to convict Xi when he is guilty but also when he is not guilty.

The point to realize is that significance is not so much about the underlying truth, which we cannot know, as about how high we set the standard for conviction. If the standard is low, we are more likely to convict when the defendant is truly guilty (we find an effect when there truly is one); we have more power. But we are also more likely to convict when he is not guilty (we find an effect when there is not truly an effect). This means there is a trade-off, and increasing the significance threshold to increase power is not a panacea.

Allocation ratio and power

Another determinant of power is the fraction of the sample that is allocated to the treatment group. Normally we assume that half the sample will be allocated to the treatment group and half to the comparison group, but this is not necessarily optimal, as we discuss below.

Why does the allocation fraction impact our level of power? If we allocate more of the sample to treatment than to comparison we will have a more accurate measure of the treatment mean than of the comparison mean. Usually this is inefficient, because there are diminishing returns to adding more sample size: increasing sample size from 10 people to 20 improves our accuracy a lot, but increasing our sample size from 1,010 to 1,020 does not make much difference. Each additional person we add makes a smaller and smaller difference in accuracy. So, all else equal, our estimate of the difference between the treatment and comparison group means is maximized if we allocated half the sample to treatment and half to control. If we have two randomization cells (one treatment and one comparison) and P is the proportion of the sample allocated to treatment, we achieve the highest power when $P = 0.5$. There are exceptions in which it is useful to create unequally sized groups; we discuss these later in this chapter.

Power with clustering

Until now we have assumed that we randomly assign individuals to either a treatment or a comparison group. But sometimes we randomize at the level of the school, clinic, or community rather than at the level of the individual. In our remedial education example, rather than randomly picking children to be part of the program, we may choose 50 schools to be part of the program and then randomly select 20 children per school to interview. What does this do to our power calculations? Randomizing at the group level reduces power for a given sample size because outcomes tend to be correlated within a group. The more correlated people are within a group, the more that randomizing at the group level reduces our power.

For example, in Sierra Leone the Agricultural Household Tracking Survey (AHTS) was designed to estimate, among other things, how much rice the average farmer grew in the past year.[9] If we had ran-

[9] Results from the Government of Sierra Leone's Agriculture Household Tracking Survey can be found at Harvard's dataverse website: http://dvn.iq.harvard.edu/dvn/.

domly picked 9,000 farmers from across the country, we would have gotten a pretty accurate estimate of rice production by smallholder farmers in Sierra Leone. We might have been tempted, for logistical reasons, to randomly pick three districts and survey 3,000 farmers from each of these three districts. But this would have given us a much less accurate picture of rice production for two reasons. Some districts (such as Koinadugu) tend to grow less rice than the other parts of the country because the land and climate there are suitable for livestock. In the year in which the AHTS was conducted, the district of Bonthe in the south was hit by a damaging flood, which suppressed the rice harvest. In other words, both long-run conditions and short-run shocks meant that rice production levels were correlated within the district. This within-district correlation made an estimate of the average harvest that was based on collecting data from a few districts much less precise for a given sample size.

The same problem occurs when we are trying to estimate the impact of a program, in this case a program to promote increased rice production. In the extreme, if we randomize one village to receive the program and one to be the comparison group, we will never be able to distinguish whether any difference we find is due to village-level factors including weather shocks or to the program. Even if we survey 100 farmers in each village, we cannot isolate the village-level shocks from the program effect. It is almost as if we had a sample of 2, not 200.

So we have to randomize over many villages, some treatment and some comparison. But how many villages do we need, and how does that number change with the number of farmers per village whom we survey? Let us think back to our power-calculation bell curves, which represent the probability of finding different effect sizes if we run the same experiment multiple times. If we randomize 600 individual farmers into treatment and comparison groups, we will get pretty similar results from our experiment each time we run it because farmers will be drawn from all over the country each time. If we randomize 15 villages into a treatment group and 15 villages into a comparison group and survey 40 farmers per village, we are likely to get different estimates of the effect size each time we run the experiment. Sometimes we will pick more villages in heavy rice-growing areas for the comparison group, sometimes the other way around. Sometimes bad weather will decrease output more in treatment villages than in comparison villages, sometimes the other way around. For a given sample size, the variance of the estimator will be larger when we randomize by village than when we randomize by individual because our sample is less diversified. Our bell curves are wider, and thus our power is lower.

How much lower is our power? That depends on how many clusters (villages) we group our sample into. The more clusters, the better. It also depends on how similar people in our clusters are to each other compared to how similar they are to the population in general. If the people in our clusters are very similar to each other (i.e., the within-cluster variance is very low), our power is much lower when we move to a group-level randomization. If the people in our clusters are just as varied as the population as a whole, moving from individual-level randomization to group-level randomization does not much change our power.

A specific example brings all these factors together. An evaluation was seeking to estimate the impact on yields of the introduction of a new high-yielding rice variety in Sierra Leone called New Rice for Africa (NERICA).[10] Seed rice for the new variety was roughly twice as expensive as traditional rice seed, so the project concluded that it would be successful only if NERICA yields were at least 20 percent higher than those of traditional varieties. The average rice yield in the area of study was 484 kilograms per hectare, so a 20 percent increase would be an increase of 97 kilograms per hectare. The standard deviation of rice yields is 295, so the standardized MDE is 97/295, which is 0.33 standard deviations. The intracluster correlation in rice yields was 0.19. If the project were to be randomized at the individual level, the sample size needed to detect a MDE of 20 percent would be 146 individuals per treatment cell and 292 in total (assuming 20 percent power, 5 percent alpha, and the same number of individuals in both groups). However, given the likelihood of spillovers within villages (if farmers were to pass the new seed on to their neighbors), it was decided that the randomization would need to be at the village level. Because roughly 25 percent of any survey costs in rural Sierra Leone are accounted for by the transport costs of reaching remote communities, it made sense to interview more than one farmer once a survey team had reached a village. With 5 enumerators on each team, it was estimated that it would be possible to survey 10 farmers per day. Thus a one-day visit to a village would produce 10 surveys. So, assum-

[10] This example is inspired by an ongoing study by Rachel Glennerster and Tavneet Suri.

ing 10 farmers interviewed per village, two treatment cells (one treatment and one comparison), an equal allocation fraction, reaching an MDE of 20 percent would require surveying 80 villages (40 treatment and 40 comparison) and 800 farmers in total.

The algebraic formula for power, as well as an algebraic explanation of the formula, is given in the next module.

MODULE 6.3 An Algebraic Exposition of the Determinants of Power

Power analysis uses power functions to calculate the sample size necessary to identify a given effect size. In the previous module we explained the intuition behind why the different ingredients of the power equation affect power. In this module we explain the relationship algebraically for both individual and group-level randomization. This module is designed for those with greater familiarity with statistics. Those with less familiarity with statistics should feel free to move directly to the next module.

Individual-level randomization

As we discuss in Module 8.2, we usually analyze the results of a randomized evaluation by running a regression of our outcome variable against a dummy variable that takes the value one for those in the treatment group and zero for those in the comparison group. The coefficient of this dummy variable gives us the difference in the means of the outcome variables for the treatment group and the comparison group. We can also include control variables in the regression, which can help us explain the outcome variable. More formally,

$$Y_i = c + \beta T + X_i \gamma + \varepsilon_i,$$

where Y_i is the outcome variable for individual *i,* β is the coefficient of the treatment dummy *T,* X_i is a set of control variables for individual *i,* and ε_i is the error term. In this framework the constant *c* is the mean of the comparison group.

In this framework, the variance ($\hat{\beta}$) of our estimate of β is given by

$$Variance\ (\hat{\beta}) = \frac{1}{P(1 - P)} \frac{\sigma^2}{N},$$

where *N* is the number of observations in the experiment; σ^2 is the variance of the error term, which is assumed to be independent and identically distributed (IID); and P is the proportion of the sample that is in the treatment group. Note that we have included control variables in our equation, and thus σ^2 is the residual variance. As discussed above, including controls in this way reduces residual variance and increases power. If α is our significance level (traditionally taken as 5 percent), the critical value $t_{\alpha/2}$ is defined as

$$\Phi(t_{\alpha/2}) = 1 - \alpha/2$$

for a two-sided test, where $\Phi(z)$ is the standard normal cumulative distribution function (CDF), which tells us the proportion of the mass under the curve to the left of a particular value (in this case $t_{\alpha/2}$). Then, as illustrated in the graphical representation in Module 6.2, we can reject the null hypothesis H_0 if $\hat{\beta}$ (the estimated effect size) falls above or below the critical value—that is, if

$$\frac{|\hat{\beta}|}{SE(\hat{\beta})} > t_{\alpha/2}.$$

Note that here we are assuming a two-sided test—that is, that we are testing whether β is either larger or smaller than H_0—so we have two hypothesis curves to which to compare the null hypothesis (Figure 6.10). Usually we have a strong reason to pick the MDE that we do on the positive effect size; in contrast, we usually don't have a particular hypothesized negative effect size that we want to be able to test for. But traditionally we test symmetrically—in other words, it is as though we were testing two hypotheses, one assuming an impact of β^* and one assuming an impact of $-\beta^*$. If we are interested in testing only whether the program has a positive effect, we can perform a one-sided test, which has more power, and we replace $t_{\alpha/2}$ with t_α.

The curve to the right in Figure 6.10 shows the distribution of effect sizes if the true effect is β^*, that is, the distribution of $\hat{\beta}$ if the true impact is β^*. The one to the left shows the distribution of estimated effect sizes if the true effect size is $-\beta^*$. (For simplicity here and in the earlier graphical representation, we provide the intuition by focusing on the positive curve. But the algebra is all for a two-sided test.) The power of the test is the percentage of the area of the $\beta = \beta^*$ curve that falls to the right of the critical value. For a given κ, this proportion can be found from the value

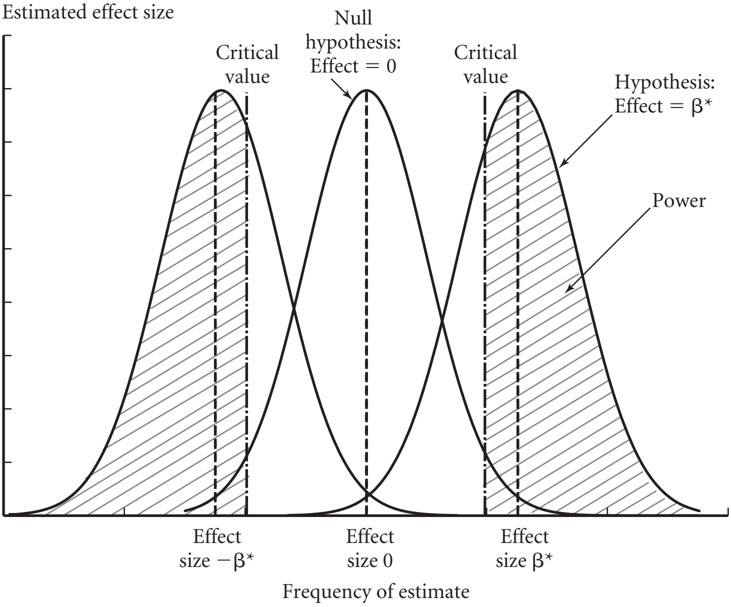


FIGURE 6.10 Distribution of effect sizes under the null and research hypotheses

$$1 - \Phi(t_{\kappa}) \equiv \Phi(t_{1-\kappa}).$$

The MDE is the minimum level of β that we can pick up with a given power. In other words, assuming that we want a power of 80 percent, the MDE is the minimum distance between the β = β* and null hypotheses for which 80 percent of the $H(\beta^*)$ hypothesis curve falls to the right of the critical value $t_{\alpha/2}$. This is given by

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

We can rearrange the equation to get

$$t_{(1-\kappa)} = \left[MDE \times \sqrt{P\,(1-P)} \times \sqrt{\frac{N}{\sigma^2}}\right] - t_{\alpha/2}.$$

Hence the power is given by

$$1 - \kappa = \Phi(t_{(1-\kappa)}) = \Phi\left(\left[MDE \times \sqrt{P(1-P)} \times \sqrt{\frac{N}{\sigma^2}}\right] - t_{\alpha/2}\right).$$

From this we can confirm the intuition represented graphically earlier in this module. A larger MDE size gives higher critical value and thus higher power (because the CDF is monotonic increasing); a larger sample size and/or smaller variance gives higher power, as does a smaller critical value. And we can see that for a given sample size, MDE, and critical level, power is maximized if $P = 0.5$.

Group-level randomization

When we randomize at the group rather than the individual level, we have to worry about two types of variation: those at the individual level and those at the group level. Our analysis equation is now

$$Y_{ij} = c + \beta T + X_i\gamma + v_j + \omega_{ij},$$

where Y_{ij} is the outcome of individual i in group j. We assume that there are J clusters of identical size n.[11] We have two error terms: one that captures shocks at the cluster level v_j, which is assumed to be IID with variance τ^2, and one that captures shocks at the individual level within a cluster, ω_{ij}, which we assume is IID with variance σ^2. The ordinary least squares estimator of β is still unbiased, and its standard error is

$$\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{n\tau^2 + \sigma^2}{nJ}}.$$

As in the case of the individual-level randomization, P is the proportion of the sample that is in the treatment group.

Comparing the standard error for the estimate of β with group-level randomization and individual-level randomization, we can see that the precision of the estimate now depends on both the within-cluster variance (τ^2) and the between-cluster variance (σ^2). The share of the total variation that is explained by cluster-level variance is given by *intracluster correlation,* ρ:

[11] If clusters are of similar but not exactly equal size, we will slightly underestimate our power if we assume that they are the same. However, given the many assumptions going into power equations, it is rarely worth it to adjust the power equation for unequal clusters. The issue of unequal-sized clusters also arises during analysis (Module 8.2).

$$\rho = \frac{\tau^2}{\tau^2 + \sigma^2}.$$

Intracluster correlation is a key component of the design effect (D), which is a measure of how much less precise our estimate is, for a given sample size, when we move from an individual-level randomization to a group- or cluster-level randomization:

$$D = \sqrt{1 + (n - 1)\rho}.$$

The design effect says that for a given sample size, the more people we survey per group, the fewer groups we have, so the less power we have. This trade-off (more people per group but fewer groups) diminishes our power most when ρ is high because the more people within groups look like each other, the less we learn from interviewing more people in the same group.

The reduction in precision of our estimated effect that comes with a move to group-level randomization has implications for sample size calculations. Specifically, Bloom shows that the MDE with J groups of size n each is given by

$$MDE = \frac{M_{J-2}}{\sqrt{P(1 - P)\,J}} \sqrt{\rho + \frac{1 - \rho}{n}}\,\sigma,$$

where $M_{J-2} = t_{\alpha/2} + t_{1-\kappa}$ for a two-sided test.[12]

This formula summarizes the relationships between power, MDE, variance, and sample size—the intuition of which we discuss in detail in the previous module. For the most part, evaluators use statistical packages to perform the calculations, but it is important to understand the intuition of how the elements relate because this allows us to design more powerful experiments, as we discuss in the next module.

Module 6.3 summary

How power relates to each of the following:

- *Significance level:* The lower the level of significance we require, the more likely we are to reject the null hypothesis and find a

[12] Howard S. Bloom, "Randomizing Groups to Evaluate Place-Based Programs," in *Learning More from Social Experiments* (New York: Russell Sage Foundation, 2005), 115–172.

statistically significant effect (i.e., the higher the power). However, we are also more likely to make false positive (Type II) errors.

- *MDE:* The larger the minimum detectable effect size, the higher the power.
- *Variance:* The lower the variance of the outcome of interest in the population, the lower the variance of the estimated effect size and the higher the power.
- *Sample size:* The larger the sample size, the lower the variance of our estimated effect (the closer it is to the true effect) and the higher the power.
- *Allocation fractions:* The more evenly the sample is distributed between treatment and comparison groups, the higher the power.
- *Level of aggregation in the unit randomized:* Individual randomization is more powerful than group-level randomization with the same sample size.
- *Intracluster correlation:* The more correlated outcomes are within groups in a group-level randomization, the less power.

MODULE 6.4 Performing Power Analysis

A power function *relates power to its determinants: (1) the level of significance, (2) the MDE size, (3) the unexplained variance of the outcome of interest, (4) the allocation fractions, (5) and the sample size. But where do we find these ingredients? In this module we first discuss how to find and calculate the ingredients necessary to calculate power. We then provide examples of statistical packages that can be used to perform power analysis.*

Ingredients for a power calculation

Desired power

Common levels of power that are used are 80 percent and 90 percent. If our power is 80 percent, in 80 percent of the experiments of this sample size conducted in this population, if there truly is an effect, we will be able to measure a statistically significant effect (i.e., we will be able to reject the null hypothesis).

MDE size

Choosing the most appropriate MDE size is probably the hardest and most important part of performing a power calculation. There is no right size; it is a matter of judgment. The true effect size of a program is a simple concept: it is the difference in the mean outcomes of the treatment and the comparison groups.

When we present the results of our experiment, we present the mean outcomes of the treatment and comparison groups and the difference between them in whatever units are most intuitive to understand. But for a power calculation we need to measure the effect size in a standard way; specifically, we measure it in standard deviations. We calculate the standardized effect size by dividing the effect size by the standard deviation of the outcome of interest:

> Standardized effect size = (mean of treatment group – mean of comparison group) / standard deviation of outcome.

We can perform this calculation only when we know the actual effect size (i.e., after we have done the experiment). What we need for a power calculation is the smallest effect size that we want to be able to pick up in the experiment. Any effect below this threshold may well be indistinguishable from zero and may well be interpreted as a program failure. Some people think of the MDE size as a prediction of the effect of the program or as a goal as to how well the program might perform. This is not a good way to think about MDE sizes. To understand why, let's return to the example of remedial education.

The first in a series of studies of Pratham's remedial education program was carried out in urban India. It found an effect size of 0.27 standard deviation (SD).[13] Subsequent evaluations looked at alternative versions of the same program—for example, testing the impact in rural locations. So in this case the researchers had a pretty good idea of what the effect size might be. Should subsequent evaluations use the effect size from the first evaluation as a guide to the MDE for power calculations?

[13] This study by Abhijit Banerjee, Shawn Cole, Esther Duflo, and Leigh Linden is summarized as Evaluation 2 in the appendix.

This strategy is risky. The effect size found in the first evaluation (0.27 SD) was large considering how cheap the program was to run. Even if the program had an impact of 0.2 SD, it would still be cost-effective. Imagine that a subsequent evaluation had chosen an MDE of 0.27 SD and the true effect size was 0.2 SD. There would have been a high probability that the subsequent evaluation would not have been able to distinguish the program effect from zero, raising concerns about whether the program was effective. Thus choosing too large an MDE might have made the evaluation too likely to fail to find an impact, even though the program in fact had an impact large enough to make it a cost-effective policy choice.

To discover the right way to think about MDE size, answer this question: "What size of effect would be so small that, for all practical purposes, it would be equivalent to zero?" There are three ways (not equally valid) to think about the right MDE size:

1. Using "standard" effect sizes
2. Comparing various MDE sizes to those of interventions with similar objectives
3. Assessing what effect size would make the program cost-effective

Let's look at each of these in detail. Usually it is worth thinking about our MDE in all three ways.

Using "standard" effect sizes There are rules of thumb about what constitutes a large effect size that are used by some researchers and appear in the literature. Based on the results of a large number of experiments on education, a view has emerged that an effect size on test scores of 0.2 SD is small but respectable, whereas an effect size of 0.4 SD is large.

Should we conclude that we can translate these rules of thumb to other sectors or even to other questions in the education sector? Maybe it is much easier to change outcomes such as school attendance than test scores and it is harder to change, for example, the height and weight of newborns than test scores. Maybe a reduction of 0.2 SD in the number of children suffering from anemia would be considered large. Even in the test score example, an effect size of 0.2 SD due to an inexpensive program in a district where few programs have been found to be effective might be hailed as a great success.

So although rules of thumb about what constitutes a large effect size are attractive because they can be taken off the shelf and used in any context, this independence of context is also its main weakness. The results of the evaluation will enter into the policymaking process as a basis for decisionmaking. As such, it seems that the optimal way of determining what is a small, medium, or large effect should come from the policy context. It is only in the policy context that even the idea of a minimally important effect size makes sense. Knowing the rules of thumb in a given sector can be a useful starting point for thinking about MDE sizes—but it should be only a starting point.

Comparing various MDE sizes to those of interventions with similar objectives Because a given program is usually one of several alternative policy options, one place to get a MDE size is from published evaluations of similar programs. By *similar programs* we mean programs that had similar goals. For example, programs that aimed to increase the number of years of schooling and have been rigorously evaluated include school input programs, scholarship programs, information programs, cash transfer programs, and health programs. Although they had very different designs and were evaluated in different countries, they had the same goals, and they all have reported effect sizes and costs. At the very least, we want our effect size to be comparable to those of other programs of similar cost. Ideally we would like to figure out what MDE size would make our program cost-effective compared to the alternatives.

For example, a school meals program intended to increase attendance can be compared to other programs that reduce the cost of education, such as programs offering school uniforms, conditional cash transfers, or deworming. What matters is that the comparable programs' costs are of the same magnitude and their stated goals are the same as those of our program.

However, if the alternatives have a very different cost per person reached than our program, then simply looking at the effect sizes of other programs is not sufficient. We need to be thinking about what MDE size would make our program cost-effective.

Assessing what effect size would make the program cost-effective Rational policymakers who want to achieve a certain goal—for example, increasing the number of children attending primary school—and have a fixed budget constraint would first invest in a program that increased school attendance by the most days for a given budget. If that program was very inexpensive and was able to be implemented throughout the country, state, or district, they might then invest in the second most cost-effective approach—and so on until their budget ran out.

If as evaluators we want to affect policy, we need to be able to distinguish whether our approach is in the range of cost-effective interventions. Given our estimate of the cost of the program, we can calculate what the effect size would need to be for the program to be cost-effective. This implies that MDE sizes will be smaller for less expensive programs because such programs are cost-effective even if they have small impacts. In other words, ironically, inexpensive programs are the most costly to evaluate because they require a large sample size for the evaluation to detect those small effects.

For cost-effectiveness comparisons, and indeed any comparisons of alternative policy options, we care not only about distinguishing our effect size from zero; we also care about having a precise estimate. Imagine that we determine that a remedial education program would be cost-effective if it achieved a test score improvement of at least 0.15 SD. If we pick 0.15 as the MDE size and 80 percent power, that means that there is an 80 percent chance that (if the effect really is 0.15) we will be able to distinguish the effect size from zero (in other words, that the confidence interval around our estimated effect size will not include zero). But we may well have a large confidence interval around our estimated effect. It may range from 0.28 to 0.02. If we are trying to compare this program with alternatives, such a wide confidence band may make it hard to make clear-cut comparisons. If we want to be able to measure the effect size precisely, we have to take this into account when setting the MDE. For example, if an effect size of 0.15 SD would make the program cost-effective, an effect size of 0.05 would mean that the program was not cost-effective, and we wanted to be able to distinguish between the two, we would need to have a MDE of 0.1.

Policymakers are influenced by many factors other than cost-effectiveness comparisons. They may, in particular, be influenced by absolute changes. If a program is inexpensive but creates only small absolute improvements, it may not receive as much support as one that is less cost-effective but produces improvements that are more visible. It may be worth asking the hypothetical questions "What size of change would be so small that it would not be considered worth it?" and "How much change would we need to see to act on the information?"

Choosing the number of clusters

As we discussed when we derived the power formula, for a given sample size we achieve the highest power by having the largest number of groups and one person per group. But sample size is typically constrained by our budget, and it is usually less expensive to interview one more person in the same group than one more person in a new group. For example, there are usually transport costs involved in traveling to a school or village. These fixed costs have to be paid regardless of the number of people we interview in the village.

Imagine that it costs $100 ($90 fixed cost + $10 per person) to interview the first person in a village and $10 per person after that. If our budget is $1,000 we can interview 1 person in each of 10 villages (with a total sample size of 10) or 11 people in each of 5 villages (for a total sample size of 55). Unless we have a very high degree of intracluster correlation, we will probably get more power from interviewing 55 people from 5 villages than 10 people from 10 villages.

There may be other practical considerations that make it more efficient to interview several people in one cluster. For example, the survey may be set up with one supervisor for a team of five enumerators. But a supervisor cannot effectively supervise if his five team members are each in a different village. Interviewing one person per village would mean that there would need to be one supervisor for every enumerator (which would be very expensive) or that enumerators would have to go to villages without a supervisor (which can lead to low-quality data).

So the marginal power from each additional person interviewed per cluster declines, but so does the marginal cost of each additional person interviewed per cluster. To maximize our power within a budget constraint, we have to trade off these two factors. In practice, costs tend to be lumpy. For example, if every enumerator can interview an average of 2 households in one day, it will not be very practical to interview 15 households per village with teams of five enumerators and one supervisor. It would be better to interview either 10 households (with each team spending one day in a village) or 20 (with each team spending two days in a village). With transport time and costs, if we try to interview 15 households, the teams will probably end up spending two days per village anyway.

Allocation fractions

Earlier we noted that in general, power is maximized when the sample size is allocated equally among all groups. More precisely, for a given sample size when there is one treatment and one comparison group, power is maximized when the allocation fraction is 0.5. However, we also noted that there are situations in which it makes sense not to distribute the sample equally between groups. In this section we examine each of these exceptions in turn.

When one budget pays for both program and evaluation

If there is one pot of funding to both run and evaluate a program and the only limits to our sample size are financial, maximizing power will involve having a larger comparison group than treatment group. This is because one additional treatment community or individual is more expensive than each additional comparison community or individual.

Imagine that we have $240,000 to both run a training program and evaluate it. It costs $1,000 per person to run the program and $100 per person to perform the evaluation. The randomization is at the individual level. Each additional person we add to the treatment group costs $1,100, while each additional comparison person costs $100. With an allocation fraction of 0.5, we could afford to have 200 people in treatment and 200 in the comparison group. When the allocation fraction is 0.5, we get the same increase in power by adding 1 treatment person as we do adding 1 comparison person. If we reduce the number of people in treatment by 1 and increase the number in the comparison group by 1, we reduce our power, but only very slightly, because we are still very close to balance between the groups. However, in this example we can reduce our treatment group by 1 and increase our comparison group by 11 within the same budget envelope. This will increase our power.

As we shift more of the sample from treatment to comparison, the power benefits diminish. At some point, reducing the treatment group by 1 and increasing the comparison group by 11 will actually reduce our power, even though it increases our sample size. This is clear if we think about the extreme case. When there is no treatment group left and only a comparison group, we have no power.

So how do we calculate the allocation fraction at which we maximize power? We can use the following formula, which says that with a set budget that must pay for both the program and evaluation, the optimal allocation fraction is equal to the square root of the ratio of the cost-per-person in the comparison group (c_c) to the cost-per-person in the treatment group (c_t):

$$(P_T/P_C) = \sqrt{(c_c/c_t)}.$$

In the example above, we maximize power with an allocation fraction of 22 percent.

If this approach sounds too complex, an alternative is to calculate different combinations of sample size and allocation fractions that we can afford with our budget and plug them into a program that calculates power (for example, in the Stata or Optimal Design software discussed below) and see which one gives the greatest power or the lowest MDE size.

When the MDE size varies by treatment group

If we have more than one treatment group, we may want to have different MDE sizes for the different treatment groups. In particular, if treatment 1 is much less expensive than treatment 2, treatment 1 would be cost-effective with a MDE size smaller than that for treatment 2. We will want to have a MDE that is at least as small as the cost-effectiveness cutoff, which suggests that we might choose a smaller MDE size for treatment 1 than for treatment 2. If we are interested only in comparing each treatment group to the comparison group, we will put more of the sample into treatment 1 than into treatment 2.

However, our decision is more complex if we are also interested in directly comparing treatment 1 with treatment 2. The appropriate MDE for a comparison between the two treatments may be even smaller than that between treatment 1 and the comparison. This is particularly true if the two treatments are variants of the same program with relatively similar costs. Even quite small differences in effect might be enough for us to conclude that one is better than the other. If we really want to be able to pick up these subtle differences between the two treatments, we will have to have large sample sizes in both treatment groups. We will discuss this conundrum more below. But the general point remains: with more than one treatment we may have more than one MDE and thus may want uneven allocation fractions.

When the comparison group plays a particularly important role

When there is more than one treatment, there is more than one pairwise comparison that can be analyzed. We can compare treatment 1 with treatment 2, treatment 1 with the comparison, and treatment 2 with the comparison. We sometimes want to compare all the treatments combined against the comparison (if, for example, the two treatments are two variants of the same program that are not very different from each other). As we mentioned above, it can be hard to distinguish one treatment from the other. If costs are similar, if one treatment is even slightly more effective than the other, it would be of policy interest. One approach is to have a large enough sample to pick up even small differences between treatments, but most experiments cannot afford the very large samples this requires. The alternative is to recognize that we are unlikely to be able to distinguish between the two effect sizes and that the core of the experiment will be comparing the different treatment groups to the comparison group. If this is the case, the comparison group will play a particularly large role.

Consider an evaluation in which the main questions of interest will involve comparing

Treatment 1 versus the comparison
Treatment 2 versus the comparison
Pooled data from treatments 1 and 2 versus the comparison

In this case, data from the comparison group are used in all three pairwise comparisons, while data from treatment group 1 are used in only two of the three. In one of the pairwise comparisons, treatment 1 data are pooled with treatment 2 data and there will be more data points in the combined treatment group than in the comparison group. Thus if we add one person to the comparison group we help improve the power of three of the pairwise comparisons; if we add one person to treatment 1 we improve the power of only two comparisons, and for one of these comparisons we already have more power than for the other cases.

For these reasons, researchers often allocate more of the sample to the comparison group than to individual treatment groups. They are particularly likely to do this if they think they are unlikely to have enough power to directly compare the different treatment groups with each other even if they have an equal allocation fraction.

There is no formula for deciding how much extra sample to put into the comparison group when it does more work, mainly because it is a matter of judgment which pairwise comparisons we care most about. A procedure that is helpful, however, is to write out all the pairwise comparisons in which we are interested. Calculate the MDE size that could be detected for the different comparisons assuming equal allocation fractions. Then rearrange the sample in a number of different ways, including placing more samples in the comparison group. Recalculate the MDE for the different comparisons with these different allocation fractions. This can help us understand where our priorities are and the trade-offs we are making in the precision with which we will answer different questions under different scenarios.

When there is greater variance in one group than another

Some treatments may have a direct effect on the variance of the outcome of interest, either increasing it or decreasing it. We might think, for example, that a program that encouraged retirees to invest their retirement savings in an annuity would reduce the variance in the income of retirees. We might therefore want to sample a larger number of people in the comparison group in order to accurately measure their highly variable income and a smaller number of people in the treatment group, whom we expect to have less variable income.

Researchers, however, very rarely use uneven allocation fractions for this reason. They may feel that they don't have enough information to judge whether variance will rise or fall because of the program, or they may decide that although they have good reason to think the program will decrease variance in one indicator it will not decrease that in other indicators of interest. Take the example of an evaluation of weather insurance. Weather insurance, if it is effective, should reduce the variance of farmers' incomes. However, researchers may be interested to see if the reduced risk from the weather changes the crops farmers plant: weather insurance may actually increase the variance of crops planted. In other words, insurance may reduce variance in one indicator while increasing it in another.

Calculating residual variance

Residual variance is the variance in the outcome variables between people (or units) that cannot be explained by the program or any control variable (e.g., gender or age) we may use in our analysis. But we perform our power calculation before our analysis. So we need to use substitutes to estimate what the residual variance will be. Specifically, we can use (1) historical data from the same or a similar population or (2) data from our own pilot survey or experiment. We usually assume that the program will not increase variance, in which case residual variance can be calculated from data collected before the program is implemented. For example, we can run a regression with our outcome variable against all the variables we plan to use as controls. Most statistical packages will then show the residual variance from the regression. This can be used for power calculations.

Historical data include data collected by the investigator in past experiments or data collected by others, including national Demographic and Household Surveys. The World Bank is a good source of survey data from which variances for particular populations can be estimated. J-PAL also posts data from previous experiments at its website (www.povertyactionlab.org). Research papers often show the variances of their samples in descriptive statistics tables. However, these will give the total variance, not the residual variance after controlling for those variables that will be used for stratification or as controls in the final analysis. A power calculation done using total variance will underestimate power or overestimate the sample size we need.

The number of repeat samples

We need to decide if we are going to conduct a baseline as well as an endline survey and whether we want to collect data between these two points as well (i.e., a *midline*). The more times we collect data on the same people, the more we will reduce the residual variance and the more power we will have. However, there are diminishing returns, and for a given budget we may find that we get more power from increasing our sample size than from adding more intermediate surveys. The increase in power from collecting more than one data point on one person depends on how correlated outcomes are for individuals over time. A person's height tends to be highly correlated over time, so conducting a baseline and an endline will allow us to reduce unexplained variance a lot. In contrast, agricultural variables such as yield tend to be less correlated over time because they are highly dependent on weather shocks, which are not constant over years. The documentation for the Optimal Design software has a useful tool for estimating the effect of multiple rounds of surveys on power.[14] It is worth noting that if we interview 100 people five times each, our sample size is 100, not 500.

We discuss other factors that affect the decision of whether to conduct a baseline survey in Chapter 5, on data collection.

Temporal correlation (for power calculations with repeat samples)

If we plan to collect data on the same person more than once—for example, with a baseline and an endline survey—and we want to undertake a power calculation that reflects this, we need to include an estimate of the *intertemporal correlation* (i.e., how correlated outcomes for individuals are over time). It can be hard to find a good estimate of temporal correlation. Even if we conduct a pilot survey, we may have only one data point for each person. Only if we run a full pilot evaluation with a baseline and an endline will we have more than one estimate, which we can then use to calculate the correlation between two data points on the same individual. The alternative is to find panel data from a similar population that has been made publically available. Panel data (which include data on the same person over time) are much less common than other types of data (Demographic and Household Surveys are usually not panel surveys, whereas data from previous randomized evaluations often are). Because it is hard to get good estimates of temporal correlation in a relevant population and researchers are nervous about overestimating temporal correlation and thus having an underpowered experiment, many researchers simply ignore the fact that they will have both baseline and endline data when they perform their power calculations. They think of this as a power buffer. In other words, they know that their power is greater than they calculated, but they do not know by how much.

Intracluster correlation (for group-level randomization)

Remember that the intracluster correlation is the percentage of the total variance in the sample that is explained by the within-cluster

[14] Jessaca Spybrook, Howard Bloom, Richard Congdon, Carolyn Hill, Andres Martinez, and Stephen Raudenbush, “6.0: Repeated Measures Trials,” *Optimal Design Plus Empirical Evidence: Documentation for the “Optimal Design” Software Version 3.0,* accessed January 3, 2013, http://hlmsoft.net/od/od-manual-20111016-v300.pdf.

variance. We can estimate the intracluster correlation by using data from other surveys of similar populations or from pilot data. Most statistical packages will be able to calculate rho from a data set that includes the outcome variable and a cluster variable (for example, village ID). Our power calculations will be sensitive to this factor, so it is worth trying to track down as good an estimate as possible.

Some tools for power calculations

Once we have the ingredients for a power calculation, how do we actually conduct it? Most statistical packages, such as Stata, have sample size functions. Free tools are also available, such as Optimal Design. Although these tools will undertake the mechanics of power for us, it is important to have a solid understanding of the principles behind them.

Stata

Stata has a command, `sampsi`, which allows users to plug in all the ingredients (including the desired level of power) to calculate the sample size needed to achieve the desired power. If the sample size is entered, `sampsi` will give the level of power. The default in Stata is that power is 90 percent, the significance level is 5 percent, the number of randomization cells is two, the allocation fraction is 0.5, the variance is the same for treatment and comparison groups, and a two-sided test is being performed.

Rather than a MDE size, Stata asks for the mean of the treatment group and the mean of the comparison group (recall that the MDE is the difference between the two). Imagine that we are performing an individual randomization of a program designed to increase test scores. We use existing data on test scores to determine that the average test score prior to the intervention is 43 percent with an SD of 5 percentage points. We decide that we want to be able to detect a 2 percent change in test scores with 80 percent power, so we enter

```
sampsi 0.43 0.45, power(0.8) sd(0.05).
```

Stata will return a sample size of 99 for the treatment group and 99 for the comparison group, assuming a 95 percent critical value (reported as an alpha of 5 percent).

If we decide we want to do the same evaluation but randomize at the class level and there are 60 children per class, we can use `sampclus`

—which is currently an add-on to Stata, so has to be downloaded the first time it is used. We first do sampsi and then adjust the sample size for the clustering by indicating either how many observations there will be per cluster or how many clusters we intend to use. We will also need to input the intracluster correlation, rho (calculated, for example, using existing data and the loneway command in Stata). In this example we would input

```
sampsi 0.43 0.45, power(0.8) sd(0.05)
```

and

```
sampclus, obsclus(60) rho(0.2).
```

The result of moving to a class-level randomization in this case is that the sample size needed to detect a change in test scores of 2 percentage points increases to 1,268 students for the treatment group and 1,268 for the comparison group, or 43 classes in total.

There are a number of options for performing more complex sample size calculations in Stata, for example, if we are collecting data more than once on the same individual or unit (in which case the extent to which outcomes are correlated over time for a given unit must be specified).

Optimal Design

An alternative software package is Optimal Design, which is specifically designed to calculate power. One advantage of Optimal Design is that it shows graphically how power increases with sample size at different levels of MDE size or intracluster correlation. This can be helpful for seeing the magnitude of some of the trade-offs we have discussed in this chapter. It is possible to see, for example, just how quickly the returns to adding more individuals per cluster diminish with different levels of intracluster correlations. Another advantage of Optimal Design is its detailed manuals, which take the user through the different types of evaluation design and which options to use. Optimal Design also allows for three levels of data collection; for example, if we are evaluating a program designed to inform voters about the records of their parliamentarians, we might collect data on individuals who are clustered into polling stations that are themselves parts of constituencies.

It is important to note that when Optimal Design reports the necessary sample size for a given MDE, the software is assuming that we have two treatment cells (i.e., one treatment and one comparison cell). If we have two treatment groups and one comparison group we need to take the reported sample size and divide by two (to get the sample size per cell) and then multiply by three (the number of randomization cells in our experiment).

Optimal Design may be downloaded free from the University of Michigan at http://sitemaker.umich.edu/group-based/optimal_design_software.

The best package for you: What you are used to

Tools will differ in their execution of power calculations. For example, one difference between Optimal Design and Stata is their treatment of binary variables in clustered designs. While Stata will prompt for rho (the level of intracluster correlation), Optimal Design avoids using rho in this context (because of concerns that rho is not accurate at the tails of the binary distribution). Our understanding is that a new version of Optimal Design is under development that will provide more accurate estimates of power for binary outcomes when the mean of the outcome variable is close to one or zero.

Finally, a warning: in practice, power calculations involve some guesswork. One way to reduce the amount of guesswork is to run a pilot on our population. This will help us get a better estimate of the mean and the variance in our population and may also help us estimate what compliance rates will be. Having better parameters will improve our calculations. But even rough estimates tell us something. So even when we don't have all the parameters we need for a full power calculation, it is still worth doing some calculations based on our best estimates of parameters.

To launch into a study without having undertaken power calculations risks wasting a massive amount of time and money. Power calculations can tell us (1) how many treatments we can have, (2) how to trade off between clusters and observations within clusters, and (3) whether the research design, at least in this dimension, is feasible or not.

Case study of calculating power

In this section we work through an example of calculating sample size for a group-level randomized evaluation of a remedial education program in Mumbai. A number of other examples and exercises on power are given at www.runningrandomizedevaluations.org.

We want to test the effectiveness of a remedial education program for third-grade students in Mumbai. We are at the early stage and want to know how many schools we will need to include in the study and on how many children in each school we will need to collect data.

We need many pieces of information for our power calculation, including the mean and variance of test scores in Mumbai. We also want to have a chance to test our data collection instrument (an academic test we have designed specifically for this evaluation). We therefore perform a pilot data collection exercise in six schools, testing all the third-grade children in these schools. On average there are 100 third-grade children in each school (no school has fewer than 80 third-grade children).

With these pilot data we calculate some basic descriptive statistics:

- The mean test score is 26 percent.
- The SD is 20.
- We also calculate the intracluster correlation as 0.17.

Note that the only use of the SD is to calculate the standardized MDE because we are doing a group-level randomization.

After looking at the impacts and costs of other education programs, we decide that we want the program to have at least a 10 percent impact on test scores—in other words, a change in the mean of test scores of 2.6 percentage points (a 10 percent change from a starting point of 26 percent). Anything less will be considered a failure. This will be our MDE size, but we have to translate it into a standardized MDE size. To do this we divide by the SD. In other words, our standardized MDE is 2.6/20 = 0.13 SD.

We are now ready to start looking at some examples of sample sizes. We set alpha to 0.05 (this gives us a significance level of 95 percent, which is standard). We start by assuming that we will test 40 children in each school. We use Optimal Design to learn that we will need 356 schools in our sample (Figure 6.11, panel A)—in other words, 178 for the treatment group and 178 for the comparison group. We decide that it will be hard for the program to work with 356 separate schools. Because this is a school setting and therefore it is relatively easy to increase the number of individuals in a group that we test, we decide to test the maximum number of children per school, which

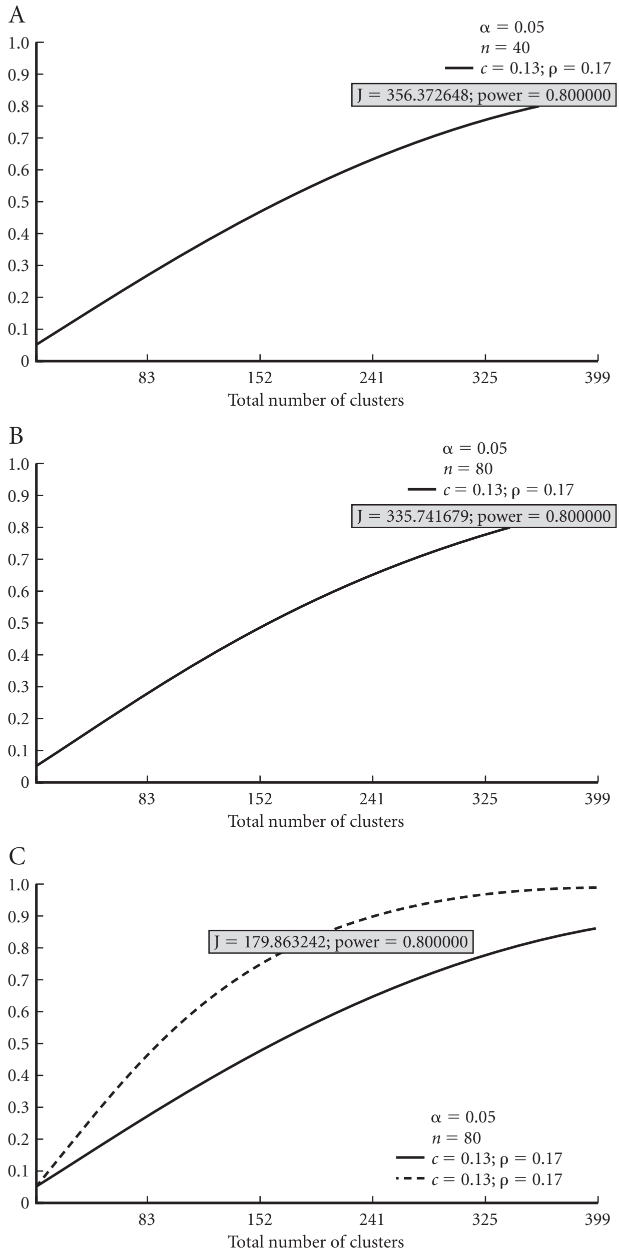


FIGURE 6.11 Number of clusters and power under different assumptions

Source: Optimal Design.

Notes: The assumptions are as follows: panel A, 40 children per school; panel B, 80 children per school; panel 3, with and without adjusting for baseline scores.

is 80. We rework the power analysis and find that for 80 percent power we will need 336 schools (Figure 6.11, panel B). In other words, doubling the number of children we test per school reduces the number of schools we will work in by only 20 out of 356, or 6 percent. Then we realize that we will make two observations on each child: before the program and after the program. Previous studies in similar populations have found that 50 percent or more of the variation in test scores in the endline survey can be explained by baseline test scores. Putting this into Optimal Design gives us a revised number of schools: 180 (Figure 6.11, panel C).

Originally we had wanted to test two different versions of the program against each other. But looking at the power analysis, we realize that we cannot afford to do this. Specifically, we had been interested in the benefit of increasing the length of the remedial sessions with children by half an hour per day. We would have considered the new version a success if it had led to a 15 percent increase in test scores compared to a 10 percent increase in test scores for the basic program. But in order to test between the two versions of the program we would have needed to be able to pick up a difference between an average test score of 28.6 (for treatment 1) and 29.9 (for treatment 2). In other words, our standardized MDE would be 0.065. The number of schools needed for this MDE (assuming that 80 children were tested per school and including baseline controls) if we were just comparing treatment and comparison groups would be 712 (356 for treatment and 356 for comparison). But we now have two treatments and a comparison group, so in fact our sample size would be 356 × 3 = 1,068 schools. This is well beyond our budget.

We therefore decide to evaluate just one treatment with 180 schools and 80 children per school for a total sample size of 14,400.

Module 6.4 summary

Where do we find the ingredients of a power calculation?

The following table summarizes the parameters we need for power calculations and where we can find them.

What we need	Where we can find it
Significance level	Conventionally set at 5 percent.
Mean of outcome in comparison group	Previous surveys conducted in the same or a similar population

What we need	Where we can find it
Residual variance in outcome of comparison group	Previous surveys conducted in the same or a similar population (the larger the variability is, the larger the sample size for a given power)
MDE size	Smallest possible effect size that would be considered important. For example, the smallest effect size that would make the program cost-effective. Drawn from discussions with policymakers and cost-effectiveness calculations.
Allocation fractions	We decide on this based on our budget constraints and prioritization of different questions. When there is only one treatment group, the allocation fraction is usually 0.5.
Intracluster correlations (when randomizing at the group level)	Previous surveys in the same or similar populations
Level of power	Traditionally, 80 percent power is used.
Number of repeat observations and intertemporal correlation	Number of times we plan to interview the same person (for example, baseline and endline). Other evaluations may have data on how correlated over time outcomes are for an individual.

MODULE 6.5 How to Design a High-Powered Study

We have to plan for power, weighing the impact of our design choices and constraints. The determinants of power suggest ways we can ensure high power at each stage of the evaluation, from design through analysis. Features of the evaluation design affect power. The most important of these are the number of treatments, the outcome measures, the randomization design, control variables, stratification, and compliance with the research design. In this module we discuss how to design and implement high-powered experiments.

When designing the evaluation

The number of questions we attempt to answer determines the number of treatment groups that we will need to have. When we calculate power, we do so for one particular pairwise comparison, say, treatment 1 versus comparison or treatment 1 versus treatment 2. Our sample size is the sample involved in that pairwise comparison. Thus if we have 180 people in the total sample and four treatments groups of equal size (45 people), our total sample size for comparing one treatment with another is 90. All things equal, the fewer the groups, the higher the power we can achieve for a given total sample.

Choose a decent sample size

A sufficiently large sample is crucial. It increases the accuracy with which we estimate impact and thus increases power. It can be a fairly expensive way of increasing power, because adding subjects increases the size of the study and, if not the number of people to be treated, at least the amount of data to be collected.

Use fewer treatment groups

Reducing the number of treatment groups increases power, but fewer groups come at a cost of fewer questions answered. As is so often the case in economics (and life), we have to make a trade-off: should we try to answer a few questions with greater precision and be reasonably confident that we will be able to detect a statistically significant effect, even if the treatment effect is quite small, or should we try to answer more questions and risk finding imprecisely estimated coefficients?

In general, it is good to have at least one question in an experiment that we can answer with a reasonably high degree of precision. For example, we may be interested in testing two different versions of a program against each other. We may not have much power to distinguish between the different versions, but we probably want to design our experiment so that it can show us whether the program as a whole has any impact with a good degree of certainty. To do this, we can combine both versions of the program into one large treatment group and compare it to the comparison group. This will give us much more power to answer this question, because the number of observations in the combined treatment group will be quite large in this case.

Randomize at the lowest level possible

The randomization level is one of the most important determinants of power. As discussed above, when randomization is done at the group level rather than person by person, the impact estimates are usually much less precise. This is because a single shock can hit an entire community, changing the outcomes of many people in that community in a correlated way. Power is much more sensitive to the number of units over which we randomize than to the number of people we survey. Often we get the same order of magnitude of power when we randomize 100 people as when we randomize 100 villages, even if the sample size in the latter case is 10 times larger (e.g., if we survey 10 people per village).

Use a design that increases compliance

Partial compliance dilutes the effect size because there is less contrast in exposure to the treatment between treatment and comparison groups. Thus the power is lower. Imagine that the true effect of a remedial education program on children who participate is 0.3 SD. But some of the children who are randomized into the treatment group fail to show up at school, and the remedial teacher in another school resigns and is not replaced. In the end, only half of the children randomized to receive the program actually complete it. The average test score difference between treatment and comparison groups is now only 0.15 SD because of lack of compliance. Because of the way sample size and effect size are related in the power equation, we need four times the sample size to pick up an effect size that is half as large.

Partial compliance can also result when some of the comparison group receives the program. In our remedial education example, imagine that there is a mix-up in the school, and some children who were randomized into the control group end up attending remedial education classes. This diminishes our measured effect size even further because it reduces the distinction between treatment and comparison groups. As above, assume that the true effect size of the program is 0.3 SD and only 50 percent of the treatment group receives the program. Now assume that 25 percent of the control group ends up receiving the program. The difference in take-up between the treatment and control groups is now only 25 percentage points; that is, the treatment group is only 25 percentage points more likely to receive the program than the comparison group. The overall effect size is only 0.3 × .025 or 0.075 SD. This is a very small effect size and will be difficult to pick up. In this case, our MDE is a quarter as large as it would have been with full compliance. This means that our sample size must now be 16 times larger.

More formally, the MDE size with noncompliance is related to the MDE size with full compliance in the following way:

$$MDE_{\text{partial compliance}} = MDE_{\text{full compliance}}/(p_t - p_c),$$

where p_t is the proportion of the treatment group that is treated and p_c is the proportion of the control group that is treated.

We can also express this in terms of sample size, namely

$$N_{\text{partial compliance}} = N_{\text{full compliance}}/(p_t - p_c)^2.$$

This suggests that a high level of compliance is vital. But sometimes take-up of a program is a critical question we want to find out about, so we don't want to be too active in encouraging participation beyond what a normal program would encourage. However, when testing the impact of a program on those who take it up, it is of critical importance to ensure that the participation rate is higher in the treatment group. As we discuss in detail in Module 8.2, it is also important to design the study to minimize the existence of defiers (individuals who take up the treatment only when they are placed in the comparison group), although in this case the problem is not so much their impact on power as the fact that they bias our estimated effect.

Encouragement designs by their nature tend to have less than full compliance because the treatment involves telling people about a program or giving them a modest incentive to participate. Encouragement designs therefore need large sample sizes. One benefit of encouragement designs, however, is that they normally involve randomization at the individual level, which does help with power.

Use a design that limits attrition

When people drop out and cannot be measured, our sample size effectively goes down, and that reduces our power. Attrition can introduce all sorts of other problems, as discussed in Module 7.2. For these reasons, and for the sake of power, choose a design that limits attrition. Ways to limit attrition are also discussed in Chapter 7.

Use stratification

As discussed in Module 4.4, when we stratify we first divide the sample into smaller groups composed of people with similar observable outcomes (such as gender or district) and then randomize from these groups into the treatment groups. This ensures that the groups are similar on the stratification variables. Stratification ensures that even in small samples, the treatment and comparison groups are balanced and have exactly equal numbers of people with different characteristics.

Stratification makes our estimate more precise for a given sample size, increasing power. By ensuring that the treatment and comparison groups are balanced on key characteristics, stratification reduces the chance that, for example, all the children with high test scores will happen to be in our treatment group and we will overestimate the treatment impact as a result. We can be sure that our estimated effect will be closer to the true effect. In terms of our power equation, balance means that there is less residual variance, which means that for a given sample size, the power is higher. In general, the more outcomes we stratify on, the better. However, it is theoretically possible to overstratify if we stratify on variables that don't have any impact on the outcome variable. We discuss the variables on which we should stratify and how to calculate residual variance with stratification in Chapter 4, on design.

Stratification is most important for small sample sizes because the law of large numbers ensures that randomization will lead to balance only in very large sample sizes.

Choose an allocation fraction

For a given sample size, power is maximized if we allocate equal shares to the treatment and comparison groups. As discussed in detail in Module 6.4, sample size may not be fixed, and there may be cases in which equal allocation fractions are not optimal.

When planning the data collection

Choose proximate outcome measures

As discussed in Chapter 5, the proximity of an outcome measure is defined by how far down the logical chain it is. In our immunization example, reducing child mortality is the ultimate objective, but it is not proximate.[15] The logical chain involves several steps: (1) camps are established and transfers provided to improve access, reliability, and incentives; (2) mothers take children to be immunized; (3) children are immunized; (4) child immunity improves; (5) children do not become sick when exposed to disease; and (6) children do not die from exposure. At each step, factors other than the program are at play:

[15] Abhijit Banerjee, Esther Duflo, Rachel Glennerster, and Dhruva Kothari, "Improving Immunisation Coverage in Rural India: Clustered Randomised Controlled Evaluation of Immunisation Campaigns with and without Incentives," *British Medical Journal* 340 (2010).

mothers attend camps for reasons other than the program, and children die for reasons other than exposure to the diseases against which the immunization protects. All these other factors become part of the variance in outcomes that cannot be explained by the program. Unexplained variance makes our estimate less precise and reduces our power. In our immunization example, we may want to use immunization rate rather than child mortality as our outcome of interest.

Some outcome measures can be too proximate to be of interest, even though they have high power. If our outcome measure is whether the camps operated, we will have power but will not have learned much about the impact of the program on people's lives. As is so often the case, a balance needs to be struck between two objectives: our desire for power and our desire to follow the logical chain as far as we can toward the ultimate impact.

Collect data on control variables

Including control variables that are strong explanitors of the outcome variable can limit the amount of unexplained variance. By reducing the amount of unexplained variance, we reduce the sample size needed to detect an effect of a given size. In other words, by reducing variance we decrease the MDE for our sample size. We discuss why this is important and how to calculate power when control variables are used below.

Collect multiple observations on one person

Individuals have idiosyncrasies that cannot be fully explained by their observable characteristics, such as their age or income. Some people are harder workers than others of a similar age and education, for example. These characteristics often persist over time: once a hard worker, always a hard worker. If we have multiple observations on the same person, such as a baseline and an endline, we can calculate a fixed effect for each person. These individual idiosyncrasies become part of the explained variance rather than the unexplained variance, which increases our power. Having a baseline also helps with stratification, because it provides variables on which we can stratify. However, it is costly to conduct additional rounds of surveys. Whether collecting more rounds of data on a small sample or expanding the sample will give us more power depends on how correlated outcomes are over time. If they are weakly correlated, a larger sample may give us higher power than a smaller sample with two survey rounds.

Plan data collection to limit attrition

Data collection strategies include collecting cell phone numbers, having a tracking module in the baseline that collects contact information on people who will know where the subject is even if he or she moves, and providing subjects with cell phones and free credit if they check in at regular intervals throughout the study.

Limit procedural variation

The fewer fluctuations there are in how a program is implemented and how the data are collected, the less variability there is in the outcome data. Some forms of measurement variation can be limited by simplifying the data collection procedures and training staff thoroughly. Limiting variability reduces the variance of our impact estimates, increasing power.

When implementing the evaluation—managing threats

Increase compliance

Strategies for increasing compliance during implementation include using a higher level of randomization between treatment groups, providing incentives to participate to all subjects, and treating only those in the cluster at the time of randomization. (For example, if we are providing school uniforms, we can give uniforms to the children already in the treatment schools at the time the program is announced. Otherwise, comparison group children may transfer to the treatment schools to obtain the uniforms.) Of course, these have to be balanced with other considerations. For example, we said above that we can increase power by using a lower level of randomization, and now we are saying that we can increase power by using a higher level of randomization. Which is it? The fact is that different effects go in different directions. When faced with this type of trade-off, it can be helpful to make some assumptions about what the crossover would be at different levels of randomization and undertake a power analysis for the different designs to see how the two factors balance out in a particular context.

Limit attrition

As discussed above, limiting attrition helps improve power. Preventing attrition needs to be a focus at all times in an evaluation, including during implementation. If the budget allows, it is useful to check in with participants (in the treatment and comparison groups) as the program is rolled out to become aware of whether participants are moving.

When undertaking the impact analysis

Use control variables

Controlling for covariates reduces unexplained variance, which increases power. Any variable we want to add as a covariate should not be influenced by the treatment, which usually means that we must use variables that were collected before the intervention started. If we collect data on covariates and use it in the analysis, this will increase the power of our tests. In our remedial education example, final test scores are correlated with lots of factors other than the program, including participants' ages, genders, and, most important, baseline test scores. If we conduct our analysis in a regression framework and put in age, gender, and baseline test scores as control variables, the unexplained variation decreases. Even if we end up having more children who are older in our treatment group than in our control group, if we use control variables we will not attribute the higher test scores associated with age to the treatment. Our estimated effect size will, on average, be closer to the true estimate (it will be more precise), and we will have more power. Control variables reduce the "background noise." This makes it easier to detect the true impact, and thus increases the power, of the experiment.

Choose a significance level

The conventional significance level is 5 percent, and it makes sense to use this when performing power analysis. Using a larger significance level may formally give us more power, but only because we have changed the rules of the game. When we write up our results, we may want to point out where there are effects, even if they are significant only at the 10 percent level, but these results will always be looked at with some hesitation. We should not plan from the outset to be able to pick up only effects with 10 percent significance.