Lecture 3: Power Calculations

Bryan S. Graham, UC - Berkeley & NBER

January 29, 2019

When conducting a program evaluation a basic question is how much data should be collected? The collection and processing of data can be costly. Particularly if the target population is scattered over a large, difficult to access, area, and in person interviews are required. A well designed evaluation therefore requires a well designed sampling and data collection plan.

An initial task is to clearly define the population of interest. The more specific the better. Examples are "out of school youth aged 13 to 18 living in urban areas", "farmers with less than 50 manzanas of land on the North Coast" or "likely voters". When evaluating an existing program, the program implementer will hopefully have a good idea of the population they are targeting for services. Sometimes the population of interest is simply "the people targeted by the organization running the program". This last case, while common in practice, may raise concerns about the generalizability of any evaluation findings.

We will return to the mechanics of sampling from the target population of interest later. For now we assume the ability to randomly sample from this population. Furthermore we can sample units – individuals, web users, households, firms – that have been exposed to the program under evaluation, as well as units who have not had access to the program. We will call the first group of units *treated*, the second *controls*. How these two groups are constituted in practice is an important question, but one we will also defer from answering for now.

How many treated and control units should we survey? To fix ideas lets assume we are interested in evaluating an employment training program for young men. The outcome of interest is earnings. The target population is young men aged 14 to 18 who are neither employed nor in school. The program begins in early 2015 and lasts for three months. We will measure the outcome interest, post-training weekly earnings in early 2016.

We decide to sample a total of N young men, measuring their earnings in early 2016. A total of N_1 of these men will have participated in the training program a year earlier, while

N_0 will serve as control units ($N = N_0 + N_1$). Let $\lambda = N_1/(N_0 + N_1)$ denote the fraction of treated individuals in our sample (i.e., the fraction that received training).

Let $\bar{Y}_0$ denote average measured earnings across the N_0 control units and $\bar{Y}_1$ average measured earnings across the N_1 treated units:

$$\bar{Y}_0 = \frac{1}{N_0}\sum_{i=1}^{N_0} Y_i,\ \bar{Y}_1 = \frac{1}{N_1}\sum_{i=N_0+1}^{N} Y_i.$$

Appealing to the Central Limit Theorem (CLT), we have (approximately)

$$\bar{Y}_0 \sim \mathcal{N}\left(\mu_0, \frac{\sigma_0^2}{N_0}\right),\ \bar{Y}_1 \sim \mathcal{N}\left(\mu_1, \frac{\sigma_1^2}{N_1}\right).$$

If training had no effect on earnings, then $\mu_0 = \mu_1$ and $\sigma_0^2 = \sigma_1^2$. We will focus on the mean effect of training. Let $\theta = \mu_1 - \mu_0$ be the mean earnings difference in 2016 across individuals who received training a year earlier versus those who did not. Our null hypothesis is no average effect. We consider the *one-sided alternative* that the training had a positive effect on earnings.

$$\begin{aligned} H_0 &: \theta \leq 0 \\ H_1 &: \theta > 0 \end{aligned}$$

We estimate the average effect on training on earning by the mean difference $\hat{\theta} = \bar{Y}_1 - \bar{Y}_0$:

$$\hat{\theta}\Big|\,\theta \sim \mathcal{N}\left(\theta, \frac{\sigma_0^2}{N_0} + \frac{\sigma_1^2}{N_1}\right)$$

Centering this statistic at the null effect of zero and standardizing by its standard error yields

$$Z = \frac{\hat{\theta}}{\sqrt{\frac{\sigma_0^2}{N_0} + \frac{\sigma_1^2}{N_1}}} = \frac{\sqrt{N}\hat{\theta}}{\sqrt{\frac{\sigma_0^2}{1-\lambda} + \frac{\sigma_1^2}{\lambda}}}$$

which is a normal random variables with unit variance:

$$Z|\,\theta \sim \mathcal{N}\left(\frac{\sqrt{N}\theta}{\sqrt{\frac{\sigma_0^2}{1-\lambda} + \frac{\sigma_1^2}{\lambda}}}, 1\right).$$

We fix the size of our test in advance at $\alpha = 0.05$. We want our testing procedure to only reject with an ex ante probability of 5 percent if, in fact, training has no mean effect on

Table 1: The logic of statistical testing

		State of the World	
		H_0	H_1
Decision	Fail to Reject	$1-\alpha$ (Correct)	$1-\beta$ (Type II Error)
	Reject	α (Type I Error)	β (Correct)

Notes: Size of test equals α. Power of test equals β.

earnings. Under the null $\theta = 0$, so that our test statistic

$$Z|\,\theta = 0 \overset{H_0}{\sim} \mathcal{N}(0,1)$$

is a standard normal random variable. We use the rule "reject H_0 if $Z \geq \Phi^{-1}(1-\alpha) = 1.645$", then our procedure will have the correct size. Here $\Phi(\cdot)$ is the cumulative distribution function of a standard normal random variable (and hence $\Phi^{-1}(\cdot)$ is the quantile function). For a $\alpha = 0.05$ one-sided test we use the $1-\alpha$ quantile of a standard normal random variable as our critical value.

By setting $\alpha = 0.05$ we ensure that we will correctly "fail to reject" the null of no mean effect at least 95 percent of the time if, in fact, training has no positive effect on subsequent earnings (See Table 1 and the left panel of Figure 1). If, in fact, $\theta = 0$, corresponding to no average effect of training, then Z, the standardized mean earnings differences across trainees and controls, should only infrequently be large. Using the normal distribution we can choose the critical value of $\Phi^{-1}(1-\alpha) = 1.645$ to limit the frequency of Type I errors (i.e., false rejections of the null).

All that remains is an analysis of the properties of our procedure *under the alternative of a positive program effect,* $\theta > 0$. Under the alternative the location of the Z distribution will be shifted to the the right (see the left-hand panel of Figure 1). We will therefore correctly reject the null more often. The question is how much more often? Recall that $1-\beta$ is used to denote the frequency of Type II errors. Type II errors correspond to failures to reject the null of no effect, when in fact there is a positive effect. The frequency of correct rejections, β, is called test *power*. We'd like power to be as large as possible (see Table 1).

For a given $\theta > 0$, the ex ante probability of (correct) rejection is

$$\begin{aligned}
\beta &= \Pr\left(Z \geq \Phi^{-1}(1-\alpha)\middle|\,\theta\right) = \Pr\left(Z - \frac{\sqrt{N}\theta}{\sqrt{\frac{\sigma_0^2}{1-\lambda}+\frac{\sigma_1^2}{\lambda}}} \geq \Phi^{-1}(1-\alpha) - \frac{\sqrt{N}\theta}{\sqrt{\frac{\sigma_0^2}{1-\lambda}+\frac{\sigma_1^2}{\lambda}}}\middle|\,\theta\right) \\
&= 1 - \Pr\left(\underbrace{Z - \frac{\sqrt{N}\theta}{\sqrt{\frac{\sigma_0^2}{1-\lambda}+\frac{\sigma_1^2}{\lambda}}}}_{\text{standard normal}} < \Phi^{-1}(1-\alpha) - \frac{\sqrt{N}\theta}{\sqrt{\frac{\sigma_0^2}{1-\lambda}+\frac{\sigma_1^2}{\lambda}}}\middle|\,\theta\right) \\
&= 1 - \Phi\left(\Phi^{-1}(1-\alpha) - \frac{\sqrt{N}\theta}{\sqrt{\frac{\sigma_0^2}{1-\lambda}+\frac{\sigma_1^2}{\lambda}}}\right). \qquad (1)
\end{aligned}$$

Equation (1) highlights several features of our test. First, the larger the effect of training (i.e., the greater θ), the greater the chance of a correct rejection. Likewise, the larger the sample size, the greater the chance of a correct rejection. Finally, the smaller the intrinsic variability in our earnings measure, the greater the chance of a correct rejection. A bigger treatment effect, more data, and less noise all make it easier to detect positive program effects.

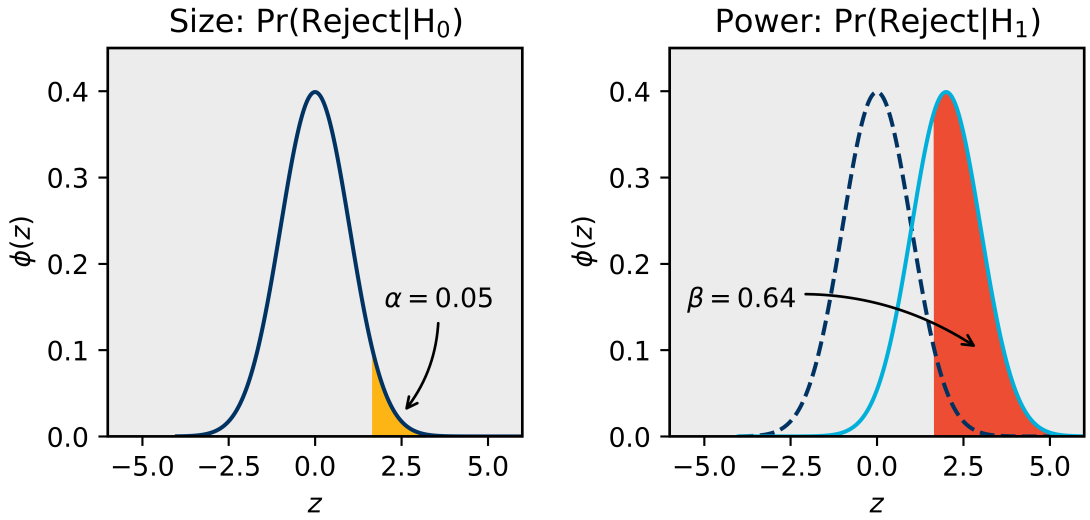
We can use (1) to answer inverse questions. For example, for a given θ, how large does our sample have to be in order to correctly reject the null with an ex ante probability of 95 percent? Solving (1) for N yields

$$N = \left(\frac{\sigma_0^2}{1-\lambda}+\frac{\sigma_1^2}{\lambda}\right)\left[\frac{\Phi^{-1}(1-\alpha)-\Phi^{-1}(1-\beta)}{\theta}\right]^2 \qquad (2)$$

Minimum sample sizes required for a pre-specified level of power

How can we turn (2) into a usable tool? Lets assume that we have information regarding earnings variability, for the target population, in the absence of training (i.e., we have a good estimate of σ_0^2). Such information can come from, for example, a regular labor force survey. Further assume that $\sigma_1^2 = \sigma_0^2$. This restriction holds under the null of "no effect of training", so it is not an unreasonable simplifying assumption. Prior to experimentation we have no way of knowing the value of σ_1^2 in any case.

Figure 1: Size and power: one-side test



Notes: The left-hand-side figure depicts the probability density function of our test statistic under the null hypothesis of no program effect. Under the null our test statistic will only exceed the critical value 100α percent of the time. By choosing αwe can control the frequency of Type I errors across repetitions of our testing procedure. The right-hand-side figure depicts the probability density function of our test statistic under the alternative hypothesis of a positive program effect. The distribution of our test statistic is shifted to the right under the alternative and therefore we will correctly reject the null at a frequency greater than size (α). The precise rejection rate is call power and can be calculated using equation (1).

If $\sigma_1^2 = \sigma_0^2$ (2) simplifies to

$$N = \frac{1}{\lambda(1-\lambda)} \left(\frac{\sigma_0}{\theta}\right)^2 \left[\Phi^{-1}(1-\alpha) - \Phi^{-1}(1-\beta)\right]^2. \tag{3}$$

If we wish to detect an average earnings effect of training as small as 100 Lempiras a week and, say, $\sigma_0 = 1,000$, then evaluating the right-hand-side of (3) yields a needed sample size of $N \approx 1,082$ for size $\alpha = 0.05$ and power $\beta = 0.95$. This calculation assumes an equal number of control and treated units (as is optimal under the given assumptions)

Sometimes researchers will discuss the *minimum detectable effect* (MDE) associated with an evaluation design.

$$\theta_{\text{mde}} = \left\{ \frac{1}{\sqrt{\lambda(1-\lambda)}} \frac{\sigma_0}{\sqrt{N}} \right\} \left[\Phi^{-1}(1-\alpha) - \Phi^{-1}(1-\beta)\right]. \tag{4}$$

The MDE is the smallest program effect that can be detected with ex ante probability β given a sample size of N. Equation (4) leads to a useful rule of thumb which you can use to assess the usefulness of evaluations that have already been conducted. Observe that the term in $\{\cdot\}$ is just the standard error of the treatment effect estimate (which should be clearly reported in any evaluation). For $\alpha = 0.05$ and $\beta = 0.8$, the term in $[\cdot]$ (equals approximately) 2.5. So if you take the reported standard error and multiply it by 2.5 you get a sense of the size of an effect that the conducted evaluation could reliably detect (if repeated many times across similar settings).

With pilot or baseline survey data, which should be collected if at all possible, it is relatively straightforward to operationalize (3) and (4). In practice researchers often deal with the absence of such data by instead considering normalized effect sizes. For example by setting θ/σ_0 equal to 0.1 or 0.2 (common choices), equation (3) can be evaluated without explicit information about the appropriate value of σ_0. Such an approach should be avoided if possible. In our example θ has a very natural metric. It may be that an effect size of 100 Lempiras a week is meaningful regardless of the level of earnings variability. But this effect will correspond to a different normalized MDE depending on the amount of earnings inequality in the target population.

A related point is that σ_0 is often partially under the control of the researcher. This is because measured earnings has two sources of variability. First there is "true" variability. Some individuals earn more or less than others. Second, there is measurement error in earnings. The degree of measurement in error is generally under (partial) control of the researcher. A better instrument, better trained enumerators, etc., can all lead to lower

measurement error and hence lower σ_0.

This observation suggests something constructive. When piloting an evaluation instrument (and I strongly encourage piloting in practice), one can test out a couple of different measures for the main outcomes of interest. For example, if interest centers on cognitive achievement, and a researcher is attempting to calibrate their measure to a specific scale, one can experiment with instruments with, say, 10 questions, 20 questions and 30 questions. Longer instruments are more difficult to field and process, but might lead to a more precise outcome measure and hence a smaller number of required respondents. With this type of pilot data in hand more informed choices about the trade-offs between sample size and more difficult to field, but possibly more precise, instruments can be made.

Adjustments for program assignment at the group level

Some programs are made available to entire groups or not. For example, a village may have a school (or clinic) or not. Furthermore, individuals who belong to the same group are generally more similar than those who belong to different groups. Let $i = 1, \ldots, N$ index groups (e.g., villages, schools, neighborhoods) and $t = 1, \ldots, T$ individuals within them. For simplicity we assume that each group has the same number of individuals (or that we sample the same number of individuals in each group). The *intra-class or intra-cluster correlation coefficient* is

$$\rho = \frac{\mathbb{C}\left(Y_{it}, Y_{it'}\right)}{\mathbb{V}\left(Y_{it}\right)}. \tag{5}$$

The intra-class coefficient provides a measure of similarity between two units who belong to the same group. For example two boys residing in the same village. Assume the outcome of interest is weight-for-height, a common measure of malnutrition. If the intra-class coefficient is high, then malnutrition is clustered in space. Intuitively hunger is systematically more common in some villages than in others. If the intra-class coefficient is low, there is very little clustering of hunger. Some children are hungry, some are not, but the mix of each type is similar across villages.

For many socioeconomic outcomes of interest the intra-class coefficient can be modest-to-high. For example outcomes across siblings tend to positively covary, as do outcomes across children in the same classroom, or households residing in the same neighborhood or village.

If the program under consideration is assigned/implemented at a "group" level it is a good idea to pilot your survey accordingly. For example you might pilot the survey to fifty individuals, five in each of 10 villages. Information from such a pilot can be used to construct a measure of ρ. Group assignment needs to be appropriately incorporated into a power

analysis.

How does group assignment affect power calculations? Assume we sample N_0 control groups, and N_1 treatment groups. In each group we interview T individuals (or firms, etc.). As before we calculate the average control and treatment outcome

$$\bar{Y}_0 = \frac{1}{N_0 T}\sum_{i=1}^{N_0}\sum_{t=1}^{T} Y_{it},\ \bar{Y}_1 = \frac{1}{N_1 T}\sum_{i=N_0+1}^{N}\sum_{t=1}^{T} Y_{it}.$$

However, since units within the same group are no long independent, we need to adjust our variances:

$$\bar{Y}_0 \sim \mathcal{N}\left(\mu_0, \frac{\sigma_0^2\left(1+(T-1)\rho_0\right)}{N_0 T}\right),\ \bar{Y}_1 \sim \mathcal{N}\left(\mu_1, \frac{\sigma_1^2\left(1+(T-1)\rho_1\right)}{N_1 T}\right).$$

To get some intuition for these variable formula, let $n = NT$ and $\lambda = N_1/N$ as before and note that

$$\begin{aligned}\frac{\sigma_0^2\left(1+(T-1)\rho_0\right)}{N_0 T} &= \frac{\sigma^2}{(1-\lambda)n} + \frac{\rho_0\sigma_0^2}{(1-\lambda)N}\frac{T-1}{T} \\ &= O\left(n^{-1}\right) + O\left(N^{-1}\right)\end{aligned}$$

so that the leading variance term is now of order N^{-1}. Because of the (positive) covariance in outcomes across groups we require observations from many difference groups to get precise estimates of program effects.

The distribution of our program effect estimate is approximately (assume there are enough groups to invoke a CLT)

$$\hat{\theta}\Big|\,\theta \sim \mathcal{N}\left(\theta, \frac{1}{NT}\left[\frac{\sigma_0^2\left(1+(T-1)\rho_0\right)}{1-\lambda} + \frac{\sigma_1^2\left(1+(T-1)\rho_1\right)}{\lambda}\right]\right).$$

If, for the purposes of power analysis, we make the simplifying assumptions $\sigma_0^2 = \sigma_1^2$ and $\rho_0 = \rho_1$ we get the simplification

$$\hat{\theta}\Big|\,\theta \sim \mathcal{N}\left(\theta, \frac{1}{N}\frac{\sigma_0^2}{\lambda(1-\lambda)}\left(\rho_0 + \frac{1-\rho_0}{T}\right)\right),$$

which, in turn, leads to the test statistic distribution

$$Z|\,\theta \sim \mathcal{N}\left(\frac{\sqrt{N}\theta}{\sqrt{\frac{\sigma_0^2}{\lambda(1-\lambda)}\left(\rho_0 + \frac{1-\rho_0}{T}\right)}}, 1\right).$$

As before we can calculate the power of our test as the probability of rejection under the alternative that $\theta > 0$. This gives

$$\beta = 1 - \Phi\left(\Phi^{-1}(1-\alpha) - \frac{\sqrt{N}\theta}{\sqrt{\frac{\sigma_0^2}{\lambda(1-\lambda)}\left(\rho_0 + \frac{1-\rho_0}{T}\right)}}\right)$$

Calculating the minimum number of groups needed to detect an effect of size θ with an *ex ante* probability of β we get

$$N = \left(\rho_0 + \frac{1-\rho_0}{T}\right)\frac{1}{\lambda(1-\lambda)}\left(\frac{\sigma_0}{\theta}\right)^2\left[\Phi^{-1}(1-\alpha) - \Phi^{-1}(1-\beta)\right]^2. \tag{6}$$

Our minimum detectable effect (MDE) is

$$\theta_{\text{mde}} = \sqrt{\rho_0 + \frac{1-\rho_0}{T}}\left\{\frac{1}{\sqrt{\lambda(1-\lambda)}}\frac{\sigma_0}{\sqrt{N}}\right\}\left[\Phi^{-1}(1-\alpha) - \Phi^{-1}(1-\beta)\right]. \tag{7}$$

Note that (6) and (7) simply scale our earlier expression, (3) and (4) above, by, respectively $\left(\rho_0 + \frac{1-\rho_0}{T}\right)$ and $\left(\rho_0 + \frac{1-\rho_0}{T}\right)^{1/2}$. The quantity $\left(\rho_0 + \frac{1-\rho_0}{T}\right)^{1/2}$ is the *design effec*t associated with the group structure of the intervention (and hence the evaluation study).[1]

The effect of large intra-class correlation on required sample sizes are be substantial. We will explore this further in the context of a worked example.

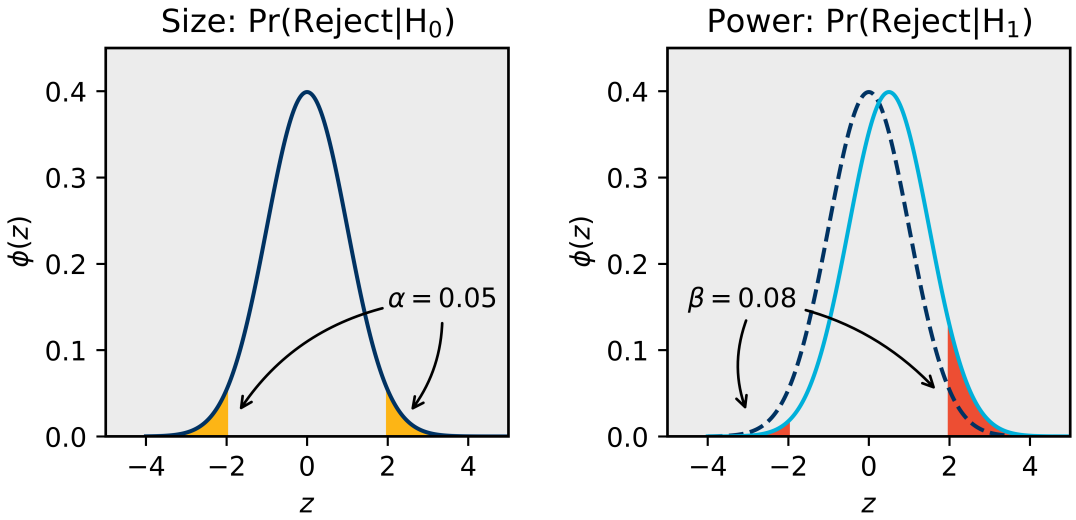
Two-sided tests

It is common practice to evaluate the null of no mean effect using a two sided test (the merits of doing so in a program evaluation context are not always clear). The basic principles and ideas used to calculate the needed sample sizes in this case remain the same, however closed-form expressions are no longer available and numerical methods are required. The StatsModel library in Python provides some basic functionality in this area, as does the 'power' command in Stata (although neither allow for non-zero intra-class coefficients; although one can often tease out the required adjustments by inputing to the program $\sigma_0\sqrt{\rho_0 + \frac{1-\rho_0}{T}}$ in place of σ_0). In practice a closed form approximation will also work in many case (see below).

To calculate the power of a two-sided test for $\theta > 0$ we evaluate (allowing for intra-cluster

[1]Formula (7) appears in Duflo et al. (2007) as their equation (12) with a typo (and different notation). It is not formally correct when two-sided tests will used, as I explain below (although in practice it is generally a very good approximation in this case).

Figure 2: Size and power: one-side test



Notes: The left-hand-side figure depicts the probability density function of our test statistic under the null hypothesis of no program effect. Under the null the absolute value of test statistic will only exceed the critical value 100α percent of the time. This will occur with the statistic being very negative about $100\frac{\alpha}{2}$ percent of the time and very positive about $100\frac{\alpha}{2}$ percent of the time. By choosing αwe can control the frequency of Type I errors across repetitions of our testing procedure. The right-hand-side figure depicts the probability density function of our test statistic under the alternative hypothesis of a positive program effect. The distribution of our test statistic is shifted to the right under the alternative and therefore we will correctly reject the null at a frequency greater than size (α). Most of the rejections will occur because our statistic is large and positive, however we will also reject sometimes due to a large (in absolute value) negative statistic (although this occurs less frequently than under the null). Both the right and left shaded regions in the right-hand-figure need to be properly accounted for when calculating the power of the two-sided test.

correlation to derive the general result)

$$\begin{aligned}
\beta &= \Pr\left(|Z| \geq \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\middle|\,\theta\right) \\
&= 1-\Pr\left(\Phi^{-1}\left(\frac{\alpha}{2}\right) < Z < \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\middle|\,\theta\right) \\
&= 1-\Pr\left(\Phi^{-1}\left(\frac{\alpha}{2}\right) - \frac{\sqrt{N}\theta}{\sqrt{\frac{\sigma_0^2}{\lambda(1-\lambda)}\left(\rho_0+\frac{1-\rho_0}{T}\right)}} < \underbrace{Z - \frac{\sqrt{N}\theta}{\sqrt{\frac{\sigma_0^2}{\lambda(1-\lambda)}\left(\rho_0+\frac{1-\rho_0}{T}\right)}}}_{\text{standard normal}} < \right. \\
&\qquad \left. \Phi^{-1}\left(1-\frac{\alpha}{2}\right) - \frac{\sqrt{N}\theta}{\sqrt{\frac{\sigma_0^2}{\lambda(1-\lambda)}\left(\rho_0+\frac{1-\rho_0}{T}\right)}}\middle|\,\theta\right) \\
&= 1-\Phi\left(\Phi^{-1}\left(1-\frac{\alpha}{2}\right) - \frac{\sqrt{N}\theta}{\sqrt{\frac{\sigma_0^2}{\lambda(1-\lambda)}\left(\rho_0+\frac{1-\rho_0}{T}\right)}}\right) \\
&\qquad +\Phi\left(\Phi^{-1}\left(\frac{\alpha}{2}\right) - \frac{\sqrt{N}\theta}{\sqrt{\frac{\sigma_0^2}{\lambda(1-\lambda)}\left(\rho_0+\frac{1-\rho_0}{T}\right)}}\right)
\end{aligned}$$

The first line to the right of the last equality corresponds to the right-hand shaded region in Figure 2, while the second to the left-hand shaded region. We cannot solve the above equation for a closed form expression for either N or θ_{mde}. However it is easy to solve for these values numerically and this is what power analysis software typically does.

In practice the second term is often very small. If it is ignored we can solve for the approximate solutions

$$N \approx \left(\rho_0+\frac{1-\rho_0}{T}\right)\frac{1}{\lambda(1-\lambda)}\left(\frac{\sigma_0}{\theta}\right)^2\left[\Phi^{-1}(1-\alpha/2)-\Phi^{-1}(1-\beta)\right]^2. \tag{8}$$

and

$$\theta_{\text{mde}} \approx \sqrt{\rho_0+\frac{1-\rho_0}{T}}\left\{\frac{1}{\sqrt{\lambda(1-\lambda)}}\frac{\sigma_0}{\sqrt{N}}\right\}\left[\Phi^{-1}(1-\alpha/2)-\Phi^{-1}(1-\beta)\right]. \tag{9}$$

The second formula appears in Duflo et al. (2007), albeit with a small typo. It is not presented as an approximation there (although it may be in the apparent source, Bloom (2005); I have not yet checked this reference). Fortunately the approximation is excellent in most practical situations (as one might expect from studying the right-hand-side panel of Figure 2).

Case study: a hypothetical lobster/conch diving intervention

Scuba diving for lobsters and conch is a common form of employment among Miskito males in Gracias a Dios. In a 2003 survey conducted by the Honduran NGO Bayan, in collaboration with the Inter-American Development Bank (IADB), about 25 percent of sampled adult Miskito males were either active or past divers. The rate of serious injury among divers, mostly due to decompression sickness, is very high. Consequently it is of interest to explore the efficacy of alternative means of income generation for men in this region.

We can use the Bayan/IADB survey to conduct a power analysis for a hypothetical intervention targeted toward reducing participation in the lobster/conch industries. This is a challenging goal, and I will refrain from offering serious thoughts about how to achieve it here, but for concreteness we can imagine the hypothetical program attempts to make agricultural work more attractive.

The survey includes information of 1,078 adult Miskito males (i.e., aged 18 and over) scattered across 60 different villages (aldeas) in the region. Of these 1,078 respondents, 286 either are active participants in lobster diving (n = 230) or former participants (n = 56). The latter group includes many individuals with serious diving related injuries (e.g., paralysis).

To compute an estimates of σ_0^2 and ρ_0 I conducted a one-way ANOVA analysis using the above data, with villages being the factors. Using the "nearly unbiased" $\hat{\omega}^2$ estimate of ρ_0 yielded 0.144. Using the more common, but also rather biased, $\hat{\eta}^2$ estimate yielded a higher value of 0.191. I include both in the power analysis to show how costly a sloppy evaluation design can be in practice. An iPython notebook replicating these calculations is available on the course webpage.

I then consider three possible effect sizes: 0.05, 0.10, 0.15. The first corresponds to a modest, but perhaps nevertheless meaningful, reduction in participation in the lobster/conch industry. The second a medium effect and the third a fairly large effect. Certainly, unless the program was very costly to implement, a reduction in participation by 15 percentage points (off a base rate of 25 percent) would have to be viewed as a success.

For brevity, consider only the bottom row of Table 2, which gives the needed sample sizes to reliably detect the larger program effect under different assumptions about intra-cluster correlation.

Say you hired an inexperienced consultant and he erroneously assumed that the intra-class correlation was zero. Or, perhaps more likely, he did power calculations "as if" the intervention operated at the individual-level, when in fact it operates at the village level. His analysis says that you need to only sample 220 individuals, evenly divided between treatment and control. If you sample twenty individuals per village, this means you only need to

Table 2: Number of villages to sample for hypothetical lobster/conch diver program

		Intra-Cluster Correlation		
		0	0.144 ($\hat{\omega}^2$)	0.191 ($\hat{\eta}^2$)
Effect Size	0.05	97 (1,940)	361 (7,220)	447 (8,940)
	0.10	25 (500)	91 (1,820)	112 (2,240)
	0.15	11 (220)	41 (820)	50 (1000)

Notes: The table reports the minimum number of villages that would need to be sampled, evenly divided among treatment and control, to reliably detect different effect sizes for a hypothetical village-level intervention aiming to reduce the rate of lobster/conch diving among Miskito males. The total number of sampled individuals is reported in parentheses below the required number of clusters. The parameters $\Pr(\text{Diver}) = 0.265$, $\sigma_0^2 = 0.195$, and $\rho_0 = 0.144$ were estimated using a sample of 1,078 adult Miskito males residing in one of 60 villages (aldeas) in Gracias a Dios, Honduras in 2003. This sample was drawn from a special fielding of the EPHPM in La Moskitia to assist in the preparation of a diagnostic report on Lobster and Conch divers in the region. Calculations are done assuming $\alpha = 0.05$, $\beta = 0.80$, and $T = 20$ adult Miskito males are sampled per village.

send enumerators to 11 villages. Unfortunately if you follow this consultant's advice you'd be disappointed with how little you learn from the evaluation.

If you hired a very experienced consultant, she would use the Bayan/BID data to construct an estimate of the intra-class correlation as we have done here. Using the preferred $\hat{\omega}^2$ estimate suggests that a sample of 41 villages, and 820 individuals, split "equally" between treatment and control would suffice. Accounting for even the modest level of intra-cluster correlation found here increases the required sample size by a factor of four. If the consultant instead used the more common, but less preferred, $\hat{\eta}^2$, estimate of intra-cluster correlation, she would suggest sample consisting of 50 villages. Which corresponds to 1,000 individuals to survey in all.

As some of you may be aware, given the difficulties of transportation in La Mosquitia, each of these three calculations suggest meaningful differences in both program and evaluation costs. This provides an example of how high the returns to some good up front analysis and thinking can be. In practice this stage of the evaluation process is often done rather haphazardly or not at all.

A good rule-of-thumb is to conduct as careful an analysis as you can, and then try to collect 10 to 20 percent more data than the power analysis suggests is needed. This provides some insurance against modest departures from baseline assumptions as well as against difficulties encountered in the field (e.g., data can't be collected in several villages due to a

road washout).

Additional considerations

The survey paper by Duflo et al. (2007) discusses many other factors that merit consideration when conducting a power analysis. One observations I'd like to make here is about the value of blocking or stratification. In our Miskito diver example consider the possibility that much of the cross-village variation in participation is driven by village distance from the coast. Boat owners typically collect their divers beachside in La Mosquitia. In such a situation one could form three groups of villages: those on the beach, those less than an hours walk, and those further away. By evenly dividing treatment and control villages in each of these three blocks, one may be able to control for a large portion of the intra-cluster correlation measured above. This can increase power considerably. The analysis of these types of designs is more complicated and we will not discuss them in detail here. They can be especially attractive for evaluating small interventions or when budgets for data collection are tight. Again, good pre-data planning can generate substantial cost savings (at least in principle).

References

Bloom, H. S. (2005). *Learning More from Social Experiments: Evolving Analytic Approaches*, chapter Randomizing groups to evaluate place-based programs, pages 115 – 172. Russell Sage Foundation, New York.

Duflo, E., Glennerster, R., and Kremer, M. (2007). *Handbook of Development Economics*, volume 4, chapter Using randomization in development economics research: a toolkit, pages 3895 – 3962. North-Holland, Amsterdam.

