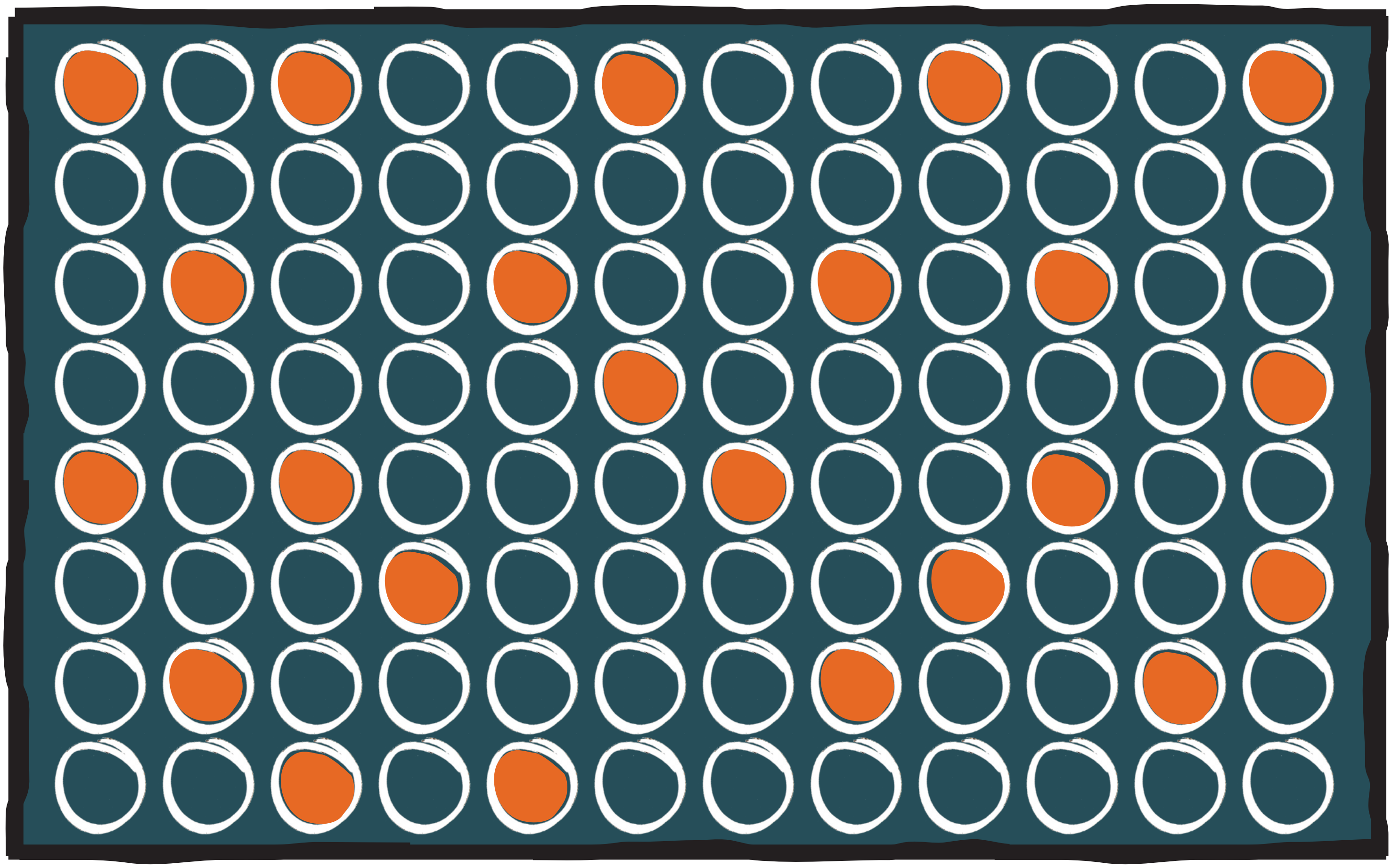
Sampling and Evaluation



A Guide to Sampling for Program Impact Evaluation

Peter M. Lance
Aiko Hattori

USAID
FROM THE AMERICAN PEOPLE

MEASURE Evaluation

Suggested citation:

Lance, P. and A. Hattori. (2016). Sampling and evaluation: A guide to sampling for program impact evaluation. Chapel Hill, North Carolina: MEASURE Evaluation, University of North Carolina.

Sampling and Evaluation

A Guide to Sampling for Program Impact Evaluation

Peter M. Lance, PhD, MEASURE Evaluation
Aiko Hattori, PhD, MEASURE Evaluation

ISBN: 978-1-943364-94-7

MEASURE Evaluation
University of North Carolina at Chapel Hill
400 Meadowmont Village Circle, 3rd Floor
Chapel Hill, NC 27517 USA
Phone: +1 919-445-9350
measure@unc.edu
www.measureevaluation.org

This publication was produced with the support of the United States Agency for International Development (USAID) under the terms of MEASURE Evaluation cooperative agreement AID-OAA-L-14-00004. MEASURE Evaluation is implemented by the Carolina Population Center, University of North Carolina at Chapel Hill in partnership with ICF International; John Snow, Inc.; Management Sciences for Health; Palladium; and Tulane University. Views expressed are not necessarily those of USAID or the United States government. MS-16-112

USAID
FROM THE AMERICAN PEOPLE

MEASURE
Evaluation

Dedicated to Anthony G. Turner

Contents

Acknowledgments

The authors of this manuscript have benefited from the comments, insights, and support of our colleagues, friends, and families. We are grateful to Paul Brodish and Lisa Parker for their extensive review of the manuscript at various stages. We also would like to acknowledge Natasha Mack, Debbie McGill, Elizabeth T. Robinson, and Kathy Doherty for their guidance with editing of the manuscript.

Chapter 1

Introduction

Program evaluation, or impact evaluation, is a way to get an accurate understanding of the extent to which a health program causes changes in the outcomes it aims to improve. Program impact studies are designed to tell us the extent to which a population's exposure to or participation in a program altered an outcome, compared to what would have happened in the absence of the program. Understanding whether a program produces intended changes allows society to focus scarce resources on those programs that most efficiently and effectively improve people's welfare and health.

The usual objective in program impact evaluation is to learn about how a population of interest is affected by the program. Programs are typically implemented in geographic areas where populations are large and beyond our resources to observe in their entirety. Therefore, we have to sample. Sampling is the process of selecting a set of observations from a population to estimate a chosen parameter — program impact, for example — for that population.

This manual explores the challenges of sampling for program impact evaluations — how to obtain a sample that is reliable for estimating impact of a program and how to obtain a sample that accurately reflects the population of interest. There are two core challenges in sampling.

First, one must select a sample that either intrinsically reflects the mix of types of individuals in the population, or one that weights (mathematically adjusts) the sample so that the mix of individuals reflects the mix in the population of interest. Impact studies usually require sampling both program participants and nonparticipants.

The second challenge is to select a sample of sufficient size to learn whatever it is that one wishes to know about impact. Larger samples contain more information and, therefore, allow us to learn more (for instance, in general, larger samples allow us to detect smaller and smaller levels of program impact).

The manual is divided into two sections: (1) basic sample selection and weighting and (2) sample size estimation. We anticipate that readers might get the most utility and comprehensive understanding from reading entire chapters rather than trying to cherry-pick portions of the discussions within them, as one might with a reference manual. This manual is more like a textbook.

Further, the manual is aimed at practitioners — in particular, those who design and implement samples for impact evaluation at their institution. Our discussions assume more than a basic understanding of sampling and some mathematical skill in applying sampling theory. That said, we are less interested in theory than in its practical application to solve sampling problems encountered in the field. We hope this manual will be a comprehensive and practical resource for that task.

Chapter 2

Basics of Sample Selection

The motivation for this manual is to discuss issues, topics, and concerns important in the context of sampling for surveys to support impact evaluations (that is, surveys intended to provide estimates of program impact). The basic motivation for a survey for impact evaluation is quite simple: to learn something about program impact within some population.

This chapter covers the basics of sample selection. This is a vital topic for impact evaluation. Without understanding it, one cannot know whether a proposed sample selection plan will yield a sample whose composition suitable for achieving the goals of an impact evaluation, particularly in terms of whether it will provide impact for the population of interest. At the same time, considerations motivated by the particulars of the selection method, such as the implications of the selection method for impact estimate precision, will inform other areas of the sampling process for impact evaluation surveys, such as sample size estimation.

Since this is a chapter in a manual on sampling and evaluation, we are technically only interested in this topic in the context of impact evaluation. Nonetheless, sample selection is a topic that transcends surveys designed to support impact evaluation and indeed is important to virtually all population survey research. Moreover, the basic story is the same for surveys designed narrowly to support impact evaluations and those intended for the more general population survey setting.

Because the topics in this chapter, while critical to sampling for impact evaluations, are rather universal, we provide a very general overview not particularly focused on impact evaluation per se. This is a deliberate decision in the interest of clarity; trying to tie this too much or too narrowly to impact evaluation would add complexity to the chapter for no real gain. In particular, we focus in our theoretical discussion and empirical simulation examples on estimation of a simple average (as opposed to a more complex impact evaluation parameter) because that is the simplest, clearest lens through which to understand the implications of sample selection choices. Nonetheless, the lessons from the simple setting of estimating an average apply as well to the context of more complicated estimation goals (it can simply be harder to show it in those cases).

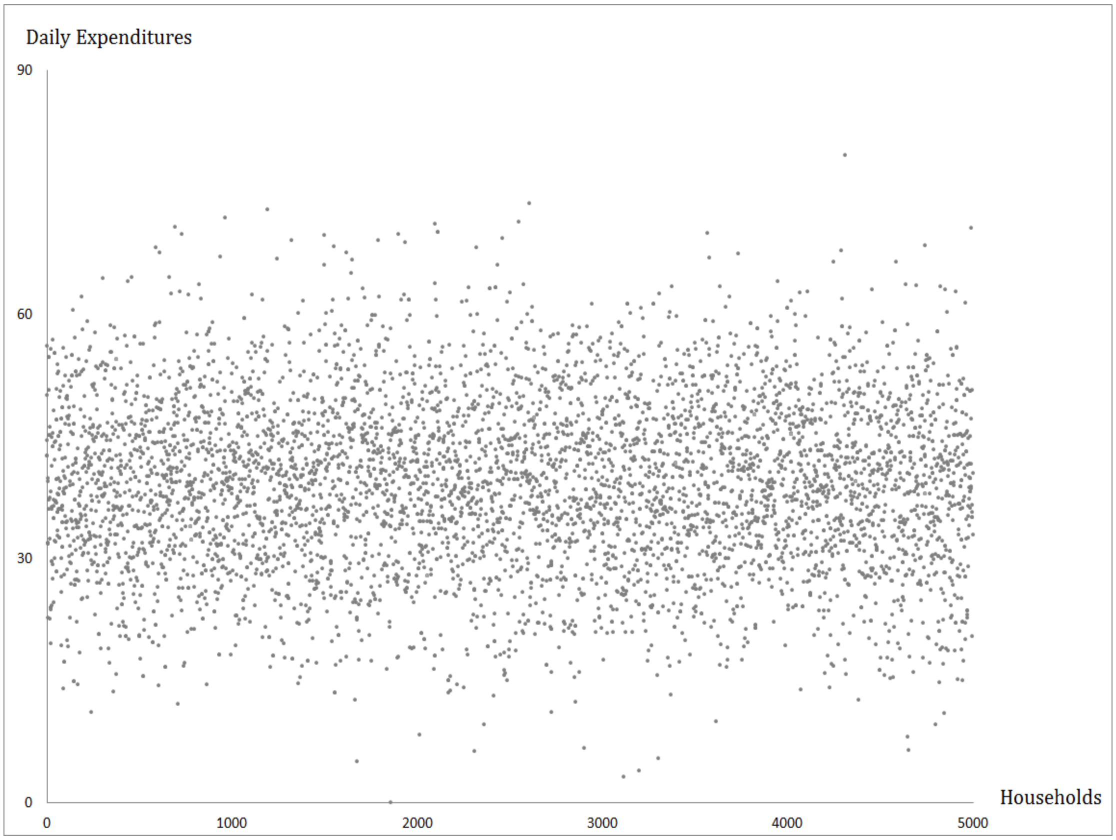


Figure 2.1. A population of 5,000 households

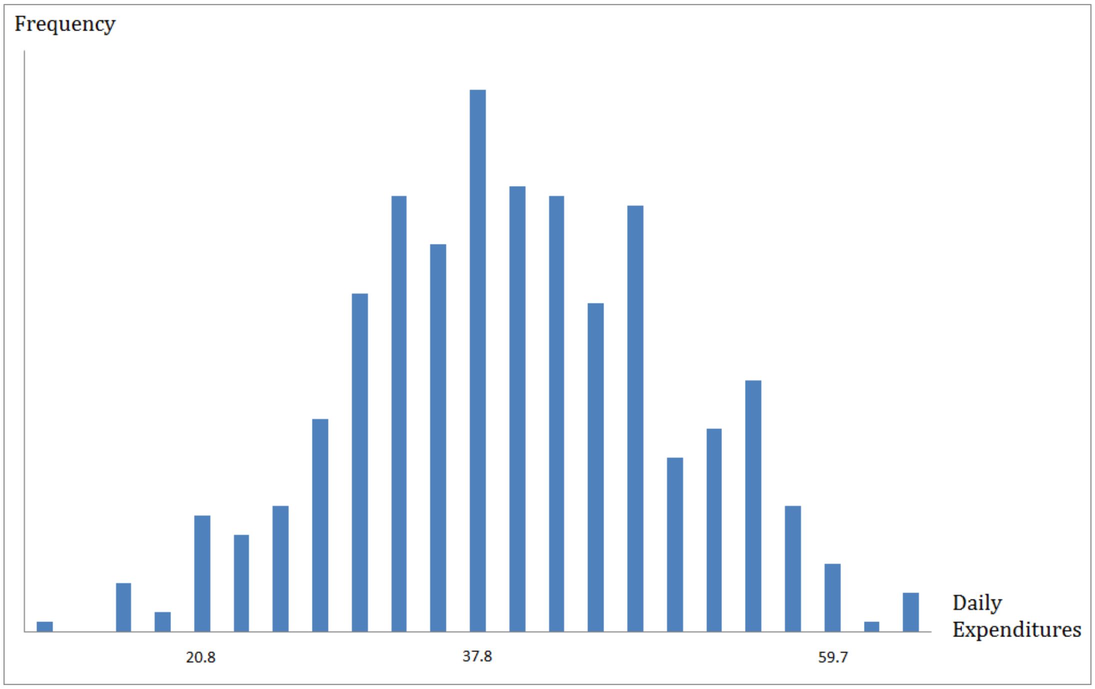


Figure 2.2. A histogram of daily expenditures

2.1 Basic Selection and Sampling Weights

We begin with our most basic goal: to learn something about a population. That *something* we refer to as a parameter. A parameter could be the rate, average, total, median, etc. of some indicator across a population. For instance, we might wish to learn something about the modern contraceptive prevalence rate across a population, such as all women ages 15–49 in Nigeria. The parameter of interest could also be the difference in two underlying parameters between two subpopulations. This is usually the focus in program impact evaluation, with the two populations (somehow) being participants and non-participants. Whatever the case, the point is that the parameter is some larger truth that we do not know about a population.

For now, we describe the parameter of interest simply as

$$\overline{Y}$$

The overline notation "$\overline{\ }$" is usually indicative of some kind of an average in statistics and econometrics. Although the parameter could be all kinds of things (a median or other quantile, something involving a higher moment, etc.) we use this notation because, in our approach to impact evaluation, we generally are thinking about some kind of an average. In particular, we tend to focus on the challenge of estimating average program impact.[1] For instance, we typically think of ourselves as trying to estimate a population parameter such as the average treatment effect, average affect of treatment on the treated, etc.

Generally, we can think about sampling goals in terms of two questions:

- What do we want to learn about $\overline{Y}$? For instance, what parameter do we want to learn about and, more to the point, what do we want to learn about it?

- For whom do we want to learn this?

The first question will be the focus of the next chapter, which deals with sample size estimation. Obviously, *what* we want to learn about $\overline{Y}$ is, at the most basic level, its value. But for the purpose of determining a sample size, we need more specifics. With what degree of precision do we want to learn about $\overline{Y}$? Or, if we want to test something about it (a test which will ultimately depend in part on our sample estimate of $\overline{Y}$), how confident do we want to be about the results of that test? These questions ultimately speak to how much information we want to have in our sample or, in other words, our sample size.

The second question speaks to our population of interest: that is, the population about which we want to learn something. In the evaluation context, we might want to learn about program impact across all adults, across the population of participants, etc. Crudely speaking, the focus of this chapter, sample selection, speaks to this second question: How do we select a sample that insures that we are focusing on the population of interest? In this chapter, we will learn basic principles of selection.

[1] Though, of course, one could consider, for instance, median program impact.

For now, let us think about the simple goal of learning something about an average across a population. To fix ideas, suppose that we wish to learn something about average daily household expenditures across some hypothetical population. By “hypothetical,” we mean in practice that this is a population we have simply made up through numerical simulations.

In this population there are 5,000 households. Figure 2.1 plots daily household expenditures per day for all 5,000 households in this hypothetical population while Figure 2.2 provides a rough histogram for daily household expenditures across the population. What these graphs provide is something almost never available (at least in contexts where we are contemplating a survey): information about every member of a population.

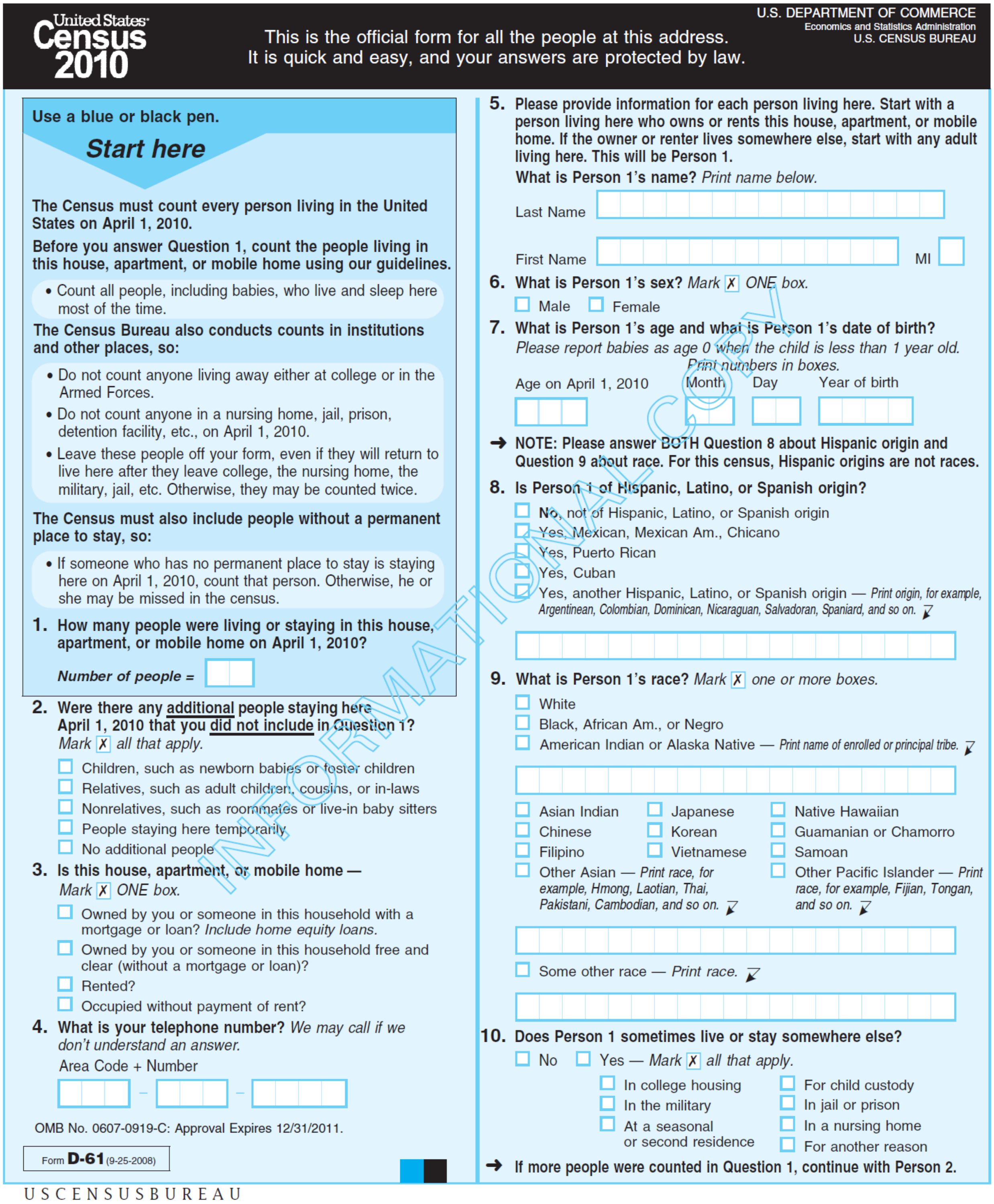
As Figure 2.1 illustrates, every household in the population of households does have *some* daily expenditure value. Moreover, as Figure 2.2 makes clear, there is some clearly discernable pattern to the distribution of daily expenditures (as opposed to, for instance, a completely random uniform distribution): There appears to be a mode at around 17–18.

We will delve into a philosophical discussion regarding the data points in Figure 2.1 and histogram in Figure 2.2 regarding whether they represent the true distribution of daily household expenditures across the population (since they offer a value for that outcome for every household in the population) or represent just one set of outcomes from some “super-population” process. There are variety of ways that the “super-population” possibility can be conceptualized, but most essentially boil down to the idea that, even if the population of households was fixed to be the same 5,000 households over time, there is a larger stable data generating process with some random component behind the daily expenditures for each household at a given observation interval. In other words, for a given observation period, some random draw for each household occurs from the larger process that establishes daily expenditures. In this sense the set of draws observed in any given period do not capture perfectly the distribution of daily expenditures, but instead offer only some information about the true data “super-population” generating process behind those draws.

For now, we abstract away from almost philosophical complications such as these, however interesting they may be. For simplicity, let us assume that the focus is on learning the actual population average for daily expenditures at the time of observation (in other words, assume more or less that we consider the population distribution and that in Figure 2.1 as to be the same thing). We want to know something about the average for this indicator across the population. The question is how we go about learning something about it.

One possibility of course would be to observe and record the daily expenditures for all 5,000 households in the population. Observing every member of the population is called a census. Censuses are sometimes a possibility for empirical research in the case of smaller populations. For larger populations, censuses are typically performed by national governments, with their far more extensive resources.

However, even then censuses have their limitations. For instance, the U.S. federal government is possessed of comparatively vast resources, and yet even in the case of the U.S. Census (conducted once every decade), the information collected per resident of the U.S. is fairly limited. Figure 2.3 shows the basic 2010 U.S. Census form that individuals filled out. As the reader can easily see, the information collected is rather modest. (There was also an additional but very basic household roster form.) Typically, more elaborate information (such as what is required to calculate daily household



United States Census 2010

This is the official form for all the people at this address.
It is quick and easy, and your answers are protected by law.

U.S. DEPARTMENT OF COMMERCE
Economics and Statistics Administration
U.S. CENSUS BUREAU

Use a blue or black pen.

Start here

The Census must count every person living in the United States on April 1, 2010.

Before you answer Question 1, count the people living in this house, apartment, or mobile home using our guidelines.

- Count all people, including babies, who live and sleep here most of the time.

The Census Bureau also conducts counts in institutions and other places, so:

- Do not count anyone living away either at college or in the Armed Forces.
- Do not count anyone in a nursing home, jail, prison, detention facility, etc., on April 1, 2010.
- Leave these people off your form, even if they will return to live here after they leave college, the nursing home, the military, jail, etc. Otherwise, they may be counted twice.

The Census must also include people without a permanent place to stay, so:

- If someone who has no permanent place to stay is staying here on April 1, 2010, count that person. Otherwise, he or she may be missed in the census.

1. How many people were living or staying in this house, apartment, or mobile home on April 1, 2010?

Number of people = □□

2. Were there any additional people staying here April 1, 2010 that you did not include in Question 1? *Mark* ☒ *all that apply.*

□ Children, such as newborn babies or foster children
□ Relatives, such as adult children, cousins, or in-laws
□ Nonrelatives, such as roommates or live-in baby sitters
□ People staying here temporarily
□ No additional people

3. Is this house, apartment, or mobile home — *Mark* ☒ *ONE box.*

□ Owned by you or someone in this household with a mortgage or loan? *Include home equity loans.*
□ Owned by you or someone in this household free and clear (without a mortgage or loan)?
□ Rented?
□ Occupied without payment of rent?

4. What is your telephone number? *We may call if we don't understand an answer.*

Area Code + Number

□□□ – □□□ – □□□□

OMB No. 0607-0919-C: Approval Expires 12/31/2011.

Form **D-61** (9-25-2008)

5. Please provide information for each person living here. Start with a person living here who owns or rents this house, apartment, or mobile home. If the owner or renter lives somewhere else, start with any adult living here. This will be Person 1.

What is Person 1's name? *Print name below.*

Last Name

First Name MI

6. What is Person 1's sex? *Mark* ☒ *ONE box.*

□ Male □ Female

7. What is Person 1's age and what is Person 1's date of birth?
Please report babies as age 0 when the child is less than 1 year old.
Print numbers in boxes.

Age on April 1, 2010 Month Day Year of birth

➜ **NOTE: Please answer BOTH Question 8 about Hispanic origin and Question 9 about race. For this census, Hispanic origins are not races.**

8. Is Person 1 of Hispanic, Latino, or Spanish origin?

□ No, not of Hispanic, Latino, or Spanish origin
□ Yes, Mexican, Mexican Am., Chicano
□ Yes, Puerto Rican
□ Yes, Cuban
□ Yes, another Hispanic, Latino, or Spanish origin — *Print origin, for example, Argentinean, Colombian, Dominican, Nicaraguan, Salvadoran, Spaniard, and so on.*

9. What is Person 1's race? *Mark* ☒ *one or more boxes.*

□ White
□ Black, African Am., or Negro
□ American Indian or Alaska Native — *Print name of enrolled or principal tribe.*

□ Asian Indian
□ Chinese
□ Filipino
□ Other Asian — *Print race, for example, Hmong, Laotian, Thai, Pakistani, Cambodian, and so on.*

□ Japanese
□ Korean
□ Vietnamese

□ Native Hawaiian
□ Guamanian or Chamorro
□ Samoan
□ Other Pacific Islander — *Print race, for example, Fijian, Tongan, and so on.*

□ Some other race — *Print race.*

10. Does Person 1 sometimes live or stay somewhere else?

□ No □ Yes — *Mark* ☒ *all that apply.*

□ In college housing
□ In the military
□ At a seasonal or second residence

□ For child custody
□ In jail or prison
□ In a nursing home
□ For another reason

➜ **If more people were counted in Question 1, continue with Person 2.**

U S C E N S U S B U R E A U

Figure 2.3. The basic 2010 U.S. Census form

expenditures accurately, which usually requires a fairly detailed expenditure instrument) is beyond the logistical and resource constraints of official national censuses, and this would presumably be the case for any other empirical study contemplating the census approach for a population of all but quite modest size.[2]

The main focus of this manual is sampling for population surveys. Population surveys take a different approach to learning something about a population parameter. Specifically, they involve selecting a sample from the population, observing the values for the behavioral indicators of interest, and using the set of observations about the sample to infer the possible population value for the parameter of interest (in this case, average household expenditures).

What this looks like in general terms is illustrated in Figure 2.4. This figure illustrates the selection of 10 households out of the population of 5,000 households (with unselected households represented in the same fashion as Figure 2.1 and selected households represented by a pronounced black dot, •). The idea behind a survey is that we would select these 10 households in essence to represent the entire 5,000 households in the population. We would then observe the daily expenditures of the selected households and, having done this for all 10 selected households, calculate average of daily expenditures across the 10.

[2] Indeed, the U.S. Census Bureau typically relies on surveys, such as the American Community Survey, to learn more detailed information about the U.S. population.

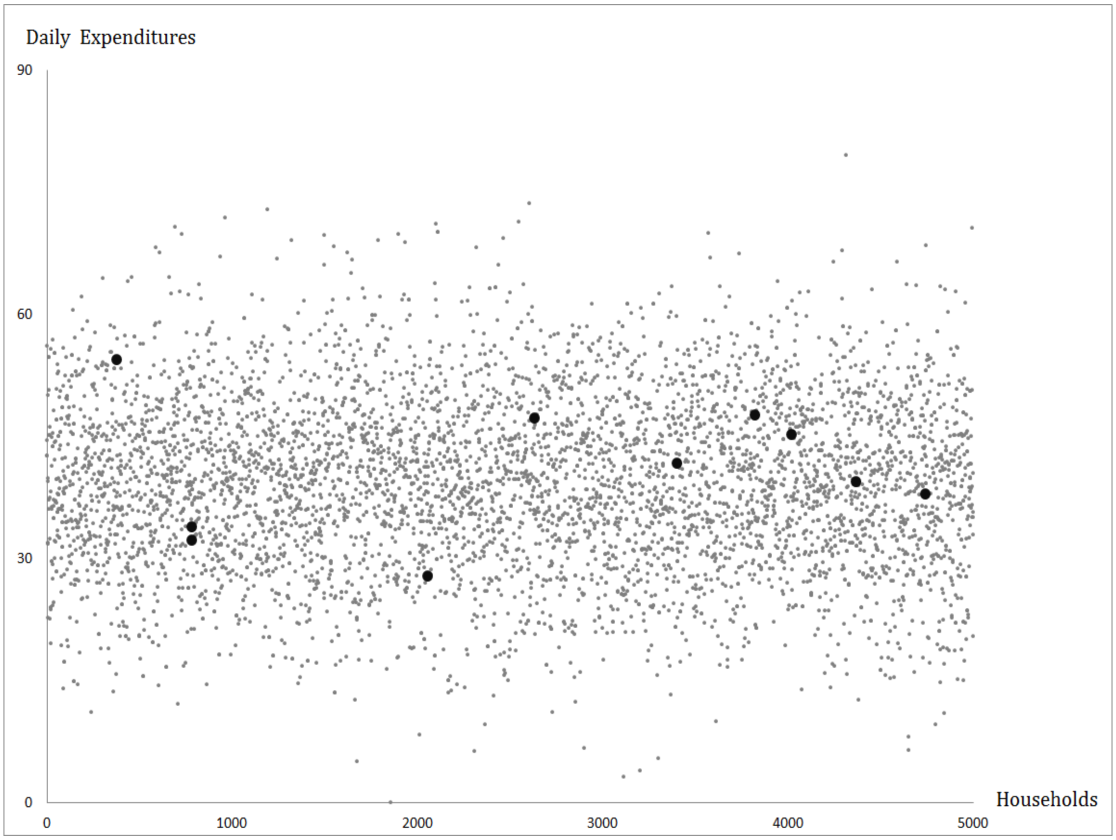


Figure 2.4. A sample of 10 households from the population of 5,000 households

This average across the 10 would serve as an estimate of the population average daily expenditures. It does not necessarily equal the population average (in fact, the probability that it will equal it *exactly* in any given realistic sample size is pretty small) but provides us with some general indication of the approximate value of the population average.

We have said that the 10 households selected for the sample represent the full population; it is worth exploring more precisely what we mean by this. We want to obtain an estimate of average daily expenditures that is indicative of the population average. To think about this, let us start to formalize our estimation approach.

The estimate of average daily expenditures for the population formed from this particular sample of 10 individuals is

$$\hat{\overline{Y}} = \frac{\sum_{i=1}^{10} Y_i}{10}$$

where Y_i is the daily expenditures of the i^{th} household of the 10 in our sample. $\hat{\overline{Y}}$ is the average for this sample of 10 households.

We still need to establish how this is an estimate. $\hat{\overline{Y}}$ is, assuming the Y_is are measured accurately, the true average for the 10 households in the sample.[3] However, these 10 households came from the population of households, and hence it seems likely that the value $\hat{\bar{y}}$ takes on for those 10 households tells us something about average daily expenditures for the larger population from which these 10 household were selected. In other words, $\bar{y}$ also estimates average daily expenditures for some larger population of households from which these 10 were selected.

But for what population does it estimate average expenditures? For instance, does it do so for the full population of 5,000 households or some subpopulation that, in the distributions of types of households within it, is not like the full population of 5,000? To begin with, suppose that the true average daily expenditures for the 5,000 households is given by $\overline{Y}$. It has some true population value (such as 39.29789). We have an estimator of average daily expenditures

$$\hat{\overline{Y}} = \frac{\sum_{i=1}^{10} Y_i}{10}$$

where $\hat{\overline{Y}}$ signifies an *estimator* of average daily expenditures. For any given sample of 10 households, it takes on a particular value that is an *estimate*.[4]

The parameter that we want the estimator $\hat{\overline{Y}}$ to estimate is called an estimand.[5] We want to estimate $\overline{Y}$. In other words, $\overline{Y}$ is our estimand. The question before us is whether $\overline{Y}$ is what $\hat{\overline{Y}}$

[3] This is true from the standpoint of the approach to population that we are taking in this conversation. If the 10 observations Y_i were simply 10 particular realizations of a a super-population process for the 10 selected households, this would not be the case. The most we could say then would be that $\overline{Y}$ is a representative estimate for those 10 households.

[4] There is often confusion around the terms "estimator" and "estimate." The estimator is in essence the rule or algorithm used to compute an estimate given a particular sample. The estimate is the particular value produced when applied to a specific sample.

[5] The term estimand is often used a bit loosely in practice. In some discussions it would seem to mean what an estimator is *actually estimating* (regardless of what we want it to estimate). At others it would seem to indicate what one *wishes to estimate* (regardless of what we are actually estimating). Though many formal definitions apparently focus on the former conceptualization, we adopt the latter definition for the present purposes.

estimates. In other words, does $\hat{\overline{Y}}$ estimate the average daily expenditures across the full population of 5,000 households, $\overline{Y}$? Or does it estimate something else (perhaps the average for some subpopulation of the 5,000)?

Suppose that we randomly selected many samples of size 10 households from the population of 5,000 households. The precise set of 10 households selected would vary across these samples and hence so would the set of observed daily household expenditures $\{Y_1, Y_2, ..., Y_{10}\}$. But this means that the estimate computed would vary from sample to sample. Moreover, since the samples were chosen randomly, this means that the estimates would vary randomly from sample to sample. The estimator $\hat{\overline{Y}}$, which produces the stream of estimates across these samples, is thus a random variable that takes on different values from sample to sample (that is, from random draw of 10 households from the population to random draw of 10 households from the population). Once we start approaching it in this way, we can start to see that what is essentially of interest is the properties of this random variable $\hat{\overline{Y}}$.

In asking whether the estimand $\overline{Y}$ is what $\hat{\overline{Y}}$ estimates, we are really asking what the central tendency of the random variable $\hat{\overline{Y}}$ is. (Or, equivalently, what is the central tendency of the distribution of the possible estimates that $\hat{\overline{Y}}$ could generate across infinite samples?) Consider, by analogy, firing a bow and arrow. We have some target (the estimand) in mind. In practice, as the vagaries of wind and atmosphere and small random mistakes in aim, arrow release, and other factors assert themselves, the arrow will almost never hit the dead center of the target. The question, however, is whether we are indeed actually shooting at the intended target. Assuming that we are, we might expect the central tendency of the arrow hits (which, in this analogy, correspond to estimates produced by successive samples) to be the center of the target. But what if we have a persistent tendency to inadvertently pull in one direction as we release, or in aiming continuously misjudge the likely flight line of the arrow to be higher than it actually turns out to be? In these cases, we would not in fact be shooting at the target and the central tendency of the arrow strikes would not be the center of that target.

Before proceeding, we digress to introduce one of the main measures of the central tendency of a random variable such as $\hat{\overline{Y}}$: the expectation $E(\cdot)$ (as in, for instance, $E(\hat{\overline{Y}})$). The expectations "operator" $E(\cdot)$ provides the expectation of a random variable.

Suppose, for instance, that X is a discrete variable that takes on three values (x_1, x_2, and x_3)

with a probability of each occurring.[6] Its expected value is then

$$E\left(X\right) = x_1 \cdot Pr\left(X = x_1\right) + x_2 \cdot Pr\left(X = x_2\right) + x_3 \cdot Pr\left(X = x_3\right)$$

where $Pr(X = x_q)$ is the probability that X takes on the value x_q. Notice that this is of the basic form

$$E\left(X\right) = \sum_{q=1}^{Q} x_q \cdot f\left(x_q\right)$$

where $\sum_q^Q$ indicates summation over the various values that X takes on (x_1, x_2, ..., x_Q) and $f(x_q)$ is the probability of outcome x_q (i.e., $Pr(X = x_q) = f(x_q)$). If X was continuous, the analog to such summation is *integration*. For a continuous random variable with probability density function $f(x)$, the expectation is thus

$$E\left(X\right) = \int_{-\infty}^{\infty} x \cdot f(x)dx$$

Therefore, the expectation is basically the weighted average of a random variable, where the weights are the probabilities of the various values that that variable can assume.[7]

[6] We digress here to discuss our variable naming conventions a bit. With the random variable X we follow the convention that specific values are represented in lowercase, as in $X = x$. In general, we do not make any distinction for the outcome of interest Y, with specific outcomes for individuals captured by Y_i, the average outcome for the population $\overline{Y}$, etc. In other words, we do not attempt some notation along the lines of $Y = y_r$ where y_r is some value in the support for Y (the "support" being the set of values for which there is a positive probability). There are two reasons for this. First, we wish to accommodate the possibility that Y is a continuous variable, for which one usually thinks of probabilities for ranges of values (as opposed to a particular value, for which the probability is typically infinitesimally small.) By contrast, X in the main text is presented as a discrete variable, whose various values in its support have a non-negligible probability of occurring. Second, and more importantly, our handling of Y is designed in part to make it consistent with our treatment of it in the potential outcomes framework, which is central to program impact evaluation and more explicitly discussed in the next chapter. We regret any confusion arising from our failure to strictly adhere to traditional statistical notation for random variables.

[7] The expectations operator has some important properties. First, if c is some constant (i.e., not a random variable, but a fixed number) then

$$E\left(c\right) = c$$

This leads to the next property, namely that

$$E\left(c \cdot X\right) = c \cdot E\left(X\right)$$

Finally, we have the additive property

$$E\left(c_1 \cdot X_1 + c_2 \cdot X_2\right) = c_1 \cdot E\left(X_1\right) + c_2 \cdot E\left(X_2\right)$$

where the cs are constants. If $c_1 = c_2 = 1$, this reduces to

$$E\left(X_1 + X_2\right) = E\left(X_1\right) + E\left(X_2\right)$$

There are, however, limits to this flexibility. First,

$$E\left(X_1 \cdot X_2\right) \neq E\left(X_1\right) \cdot E\left(X_2\right)$$

unless X_1 and X_2 are independent. Furthermore,

$$E\left(\frac{X_1}{X_2}\right) \neq \frac{E\left(X_1\right)}{E\left(X_2\right)}$$

and

$$E\left[f\left(X\right)\right] \neq f\left(E\left[X\right]\right)$$

In statistics there are two principle approaches to thinking about whether the "central tendency" of the stream of estimates produced by an estimator such as $\hat{\beta}_y$ is indeed the estimand (in this case β_y):

1. **Unbiasedness:** Because it is a random variable, the estimator $\hat{\overline{Y}}$ will have an expectation $E(\hat{\overline{Y}})$. In general, an estimator is unbiased if its expectation $E(\cdot)$ is in fact the true value of the parameter we wish to estimate (i.e., the estimand). Per the fancy terminology above, an estimator is unbiased if its expected value is indeed the estimand. Therefore, in our discussion the estimator $\hat{\overline{Y}}$ is unbiased if $E(\hat{\overline{Y}}) = \overline{Y}$. What does this mean in plain terms? An easy way to think about this is to conceptualize drawing many, many samples of size n (so far we have focused on $n = 10$) and forming the estimate $\hat{\overline{Y}}$ for each sample. If the average of these estimates across these many samples is $\overline{Y}$, then $\hat{\overline{Y}}$ is an unbiased estimator of $\overline{Y}$. For this reason, unbiasedness is sometimes referred to as being *right on average*.

2. **Consistency:** Because an estimator is a random variable, it has a probability distribution. In other words, it has probabilities attached to the possible values that it can take in a given sample. An estimator is consistent if, as sample size increases, the values that it can take on with positive probability become increasingly concentrated on the true value of the parameter one wishes to estimate. In the case of an estimator with continuous range, it is consistent if its probability density collapses around the parameter we wish to estimate as sample size grows. A somewhat crude but useful way of thinking about consistency is that it establishes that values for the estimator that deviate from that of the true parameter we wish to estimate become increasingly improbable as sample size increases.

When an estimator is said to estimate the estimand, it is usually in one or both of these senses.

Before proceeding, we illustrate these two concepts in Figures 2.5 and 2.6. In both figures, our objective is to estimate some population parameter μ^*. In Figure 2.5 we illustrate unbiasedness. In particular, we illustrate the distribution of possible estimate values for an estimator $\hat{\mu}$ applied to samples of various sizes. Regardless of size, the distribution of the possible values for the estimates is centered on μ^* with equal probability weight on either side and, as illustrated, the estimator $\hat{\mu}$ is unbiased (i.e., $E(\hat{\mu}) = \mu$). This is true regardless of the sample size — unbiasedness does not depend on sample size. Figure 2.6 illustrates consistency. In this case the central tendency of the estimates associated with the estimator $\hat{\mu}$ is not centered on μ^* at the smaller sample sizes. However, as the sample size grows larger and larger the distribution of $\hat{\mu}$ begins to collapse around the true population value μ^*.

For the present purposes we focus on unbiasedness. Given our objective (to estimate the true average daily expenditures across the population of 5,000 households) our estimator $\hat{\overline{Y}}$ would be unbiased if

$$E\,(\hat{\overline{Y}}) = \overline{Y}$$

A clear exception to the last inequality would be the case where $f(\cdot)$ is linear (e.g., $f(X) = a + b \cdot X$, where a and b are constants).

Let's delve into this a bit:

$$E\,(\hat{\overline{Y}}) = E\left(\frac{\sum_{i=1}^{10} Y_i}{10}\right)$$

By the properties of expectations operators (namely, that $E(c_1 \cdot X_1 + c_2 \cdot X_2) = c_1 \cdot E(X_1) + c_2 \cdot E(X_2)$) we can write

$$E\left(\hat{\overline{Y}}\right) = E\left(\frac{\sum_{i=1}^{10} Y_i}{10}\right) = \frac{1}{10} \cdot \sum_{i=1}^{10} E\,(Y_i)$$

It is now clear that unbiasedness hinges on the following condition

$$\frac{1}{10} \cdot \sum_{i=1}^{10} E\,(Y_i) = \overline{Y}$$

The question is what would make this condition hold.

Suppose that the expectation of Y_i is not evolving across the 10 selections (in other words, $E(Y_i)$ is the same for all $i = 1, ..., 10$). This simply means that, however we select, we are not selecting in such a fashion that we are by our design changing the (expected) type of household selected

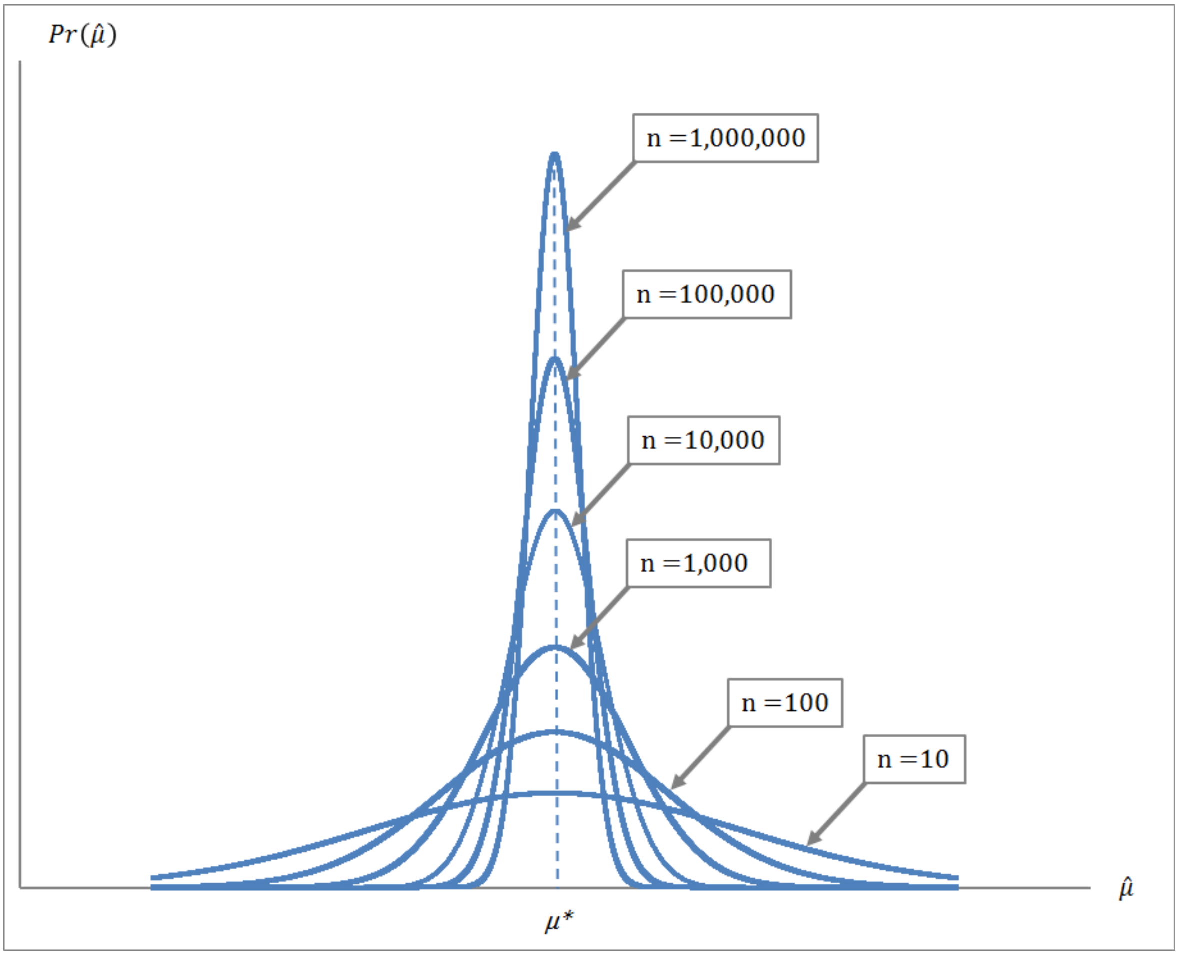


Figure 2.5. Unbiasedness: A graphical illustration

with each successive selection. This is not the same as assuming that the households we select have the same characteristics as the overall population (for instance, we could consistently select only households with more than 4 members). This simply means that $E(Y_i)$ equals some constant ω

$$\frac{1}{10}\cdot\sum_{i=1}^{10}E\left(Y_i\right)=\frac{1}{10}\cdot\sum_{i=1}^{10}\omega=\frac{1}{10}\cdot 10\cdot\omega=\omega$$

for $i=1,...,10$. But, since $E(Y_i)=\omega$ for all $i=1,...,10$, unbiasedness boils down to $E(Y_i)=\omega=\overline{Y}$.

This is a really profound result and one that formalizes (at least in this simple example) what "representative" means: It means that an observation selected for the sample is not expected *ex-ante*[8] to be different on average than the typical member of the population. Let's think about this mathematically. Suppose that, for our sample of 10,

$$Y_i=\overline{Y}+\lambda_y+\epsilon_i$$

[8] Where by "*ex-ante*" we mean, in this context, "before selection".

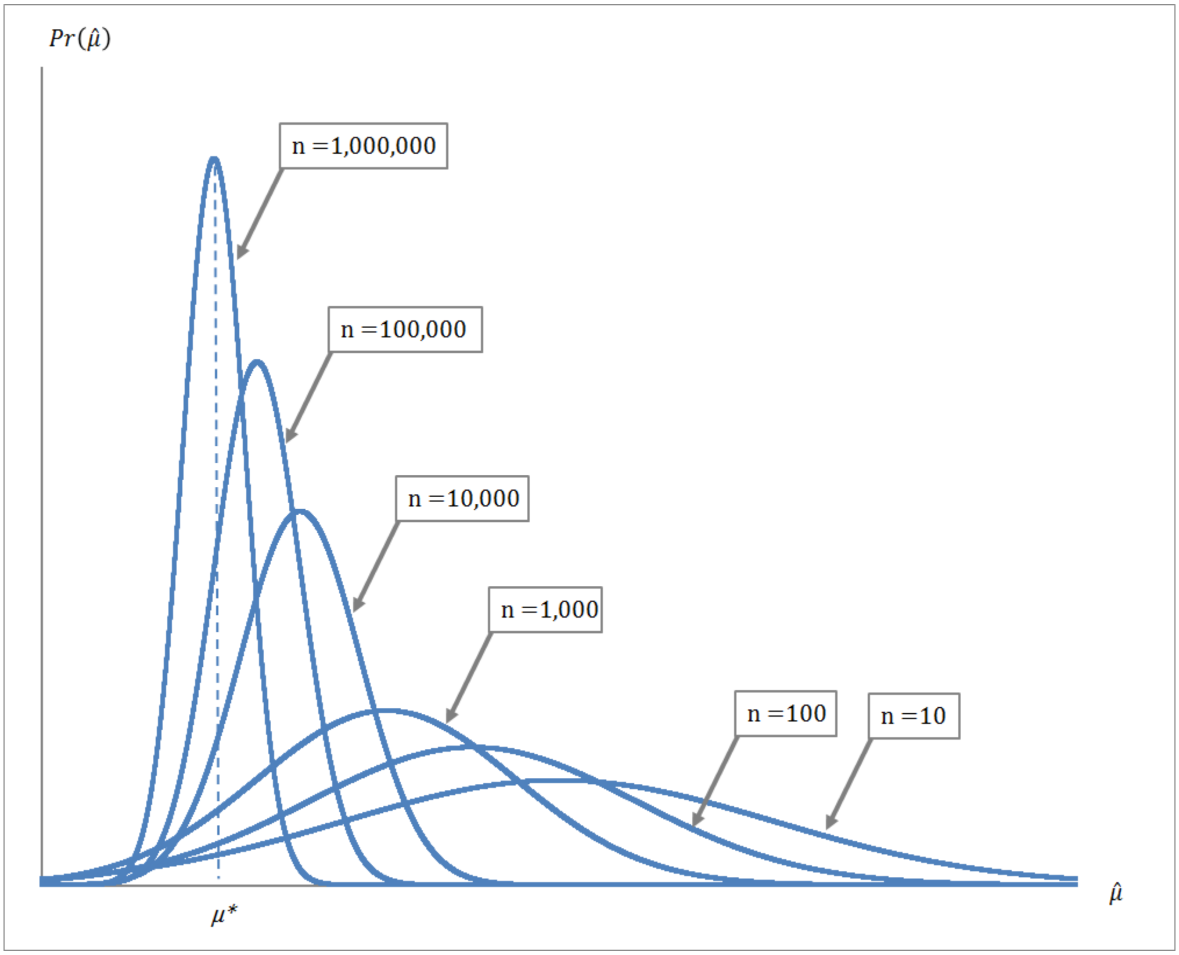


Figure 2.6. Consistency: A graphical illustration

In other words, a selected individual household's expenditure equals the population average value plus some persistent (i.e., constant) difference λ and a random difference ϵ_i. Then,

$$E(Y_i) = E\left(\overline{Y} + \lambda_y + \epsilon_i\right) = E\left(\overline{Y}\right) + E(\lambda_y) + E(\epsilon_i) = \overline{Y} + \lambda_y + E(\epsilon_i)$$

which will equal $\overline{Y}$ (the condition for unbiasedness) only if $\lambda_y = 0$ (i.e., there is no fixed or persistent wedge between expected sample and population average daily expenditures) and $E(\epsilon_i) = 0$ (any random differences between household i's daily expenditures and the same across the population are expected to equal 0).[9] In the context of this example, this means that in terms of *ex-ante* expectations, each of the households selected is not different, on average, from a typical household in the population.

In practice, of course, households *are* different — some are bigger or smaller, richer or poorer, better or less educated, better or less well exposed to markets, etc. than the "average" household across the population. All of these differences can have implications for their expected daily household expenditures. But, as of when they are selected, we have no expectation that any given household will differ from the population average. Put slightly differently, we have no expectation that it differs in type (defined in terms of the sorts of systematic background characteristics we just mentioned and thus, ultimately, in terms of daily expenditures) from the average household type in the population.

The way we can be sure of this is by making sure that we are not selecting in a fashion that would lead us to expect any type of household in the sample to be over-represented compared with their share in the overall population. The most persuasive way to insure this is to select randomly in such a fashion that no individual household in the population has a greater *ex-ante* probability of selection into the sample than any other individual household in the population.[10]

In general, to insure representativeness, random sample selection is required. When each sampling unit (in this case households) in the population of sampling units has the same probability of selection, the result is a self-weighting sample.[11] However, what is really crucial is the random part of "random selection"; as we will see below, uneven selection probabilities can be remedied through appropriate sampling weights.

We now engage in a simulation exercise to illustrate some larger truths of sample selection. The simulation process behind the selection of 10 households in Figure 2.3 was very simple:

1. Randomly simulate a population of 5,000 households. (The simulated values for daily expenditures are draws from the standard normal distribution, and then adjusted in terms of minimum value and variance to insure that the resulting distribution looks sufficiently cool

[9] Actually, the two conditions are completely distinct.

[10] As we will see later, sometimes the selection procedure involves unequal selection probabilities among households, but this can be addressed with probability weights that correct for this imbalance.

[11] Self-weighting samples are ones for which we expect that, on average, the representation of different types of population members reflects the actual share of those types in the populations. Across many repeated self-weighted samples, we would expect the frequency with which each member of the population is selected to equal the share of the population selected for each sample (for example, if 5 percent of the population was chosen for each sample, we would expect the average member of the population to be chosen for 5 percent of the repeated samples).

visibly.)

2. Draw a value from the uniform distribution for each of the 5,000 simulated households.

3. Select the 10 households with the smallest[12] values for their draws from the uniform distribution.[13]

Let us start by revealing the truth across this simulated population of 5,000 households. The true (at least as we have defined it[14]) average of daily household expenditures across this population is 39.29789. The standard deviation across the population of 5,000 households is 10.21057 with a standard error of 0.1443993.[15]

To see how the estimate of the population mean from a random sample of 10 from the population varies from sample to sample, we could conduct a numerical experiment whereby we draw independently 1,000 samples of size 10 from the population of 5,000 (for those who wish to play with this simulation exercise, the code is available in STATA do file Chap2-1.do). By "independently" we mean that selection is done independently in each of the 1,000 draws of 10 observations so that the households selected in any one sample in no way depend on those selected in any other sample. The simulation is quite simple. After simulating the population of 5,000 households we repeat the following steps 1,000 times:

1. Draw a value from the uniform distribution for each of the 5,000 simulated households.

2. Select the 10 households with the smallest draws from the uniform distribution.

3. Record the average daily expenditures for that sample of 10 households.

What we thus generated were 1,000 estimates of the population average daily expenditures from 1,000 samples of 10 households each.

[12] Or largest, or whatever the case. What is important is that the selection depend only on the draw from the uniform distribution.

[13] This is just a very simple method of random sample selection. We will discuss the issue of sample selection in more detail in a section to follow.

[14] Recall, for instance, the philosophical questions about what "population" really means.

[15] The distinction between the standard deviation and the standard error can sometimes be a source of confusion. The standard deviation measures variability across a sample. It is

$$SD = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$$

where n is the size of the sample, $\{x_1, x_2, ..., x_n\}$ is the set of observations for a variable for that sample, and $\overline{x}$ is the sample average across those n observations. This simply tells us how much dispersion a variable displays across a series of observations. In other words, it is a *feature of the sample.*

By contrast, the standard error is a measure (or, more precisely, an estimate) of the precision of the estimate of a statistic, such as the sample mean. Mathematically, the standard error is simply the standard deviation divided by the square root of the number of observations:

$$SE = \frac{SD}{\sqrt{n}}$$

The distinction between these two is even hinted at in their names. The standard "deviation" is a reference to how much the observations deviate or vary around their central tendency in a sample. The standard "error" refers to the error or variation in estimates of a statistic (such as the sample mean) from sample to sample around its true population value.

In Table 2.1 we present the estimates of population average daily expenditures from just 10 of those samples (randomly chosen from the 1,000 samples simulated). Recall that the true average daily expenditures across the population is 39.29789. As evident in Table 2.1, the estimates of average daily expenditures vary from sample to sample across these 10 example samples. This is visually illustrated in Figure 2.7, where the blue dots represent estimates of the average daily expenditures for each of the 10 observations in Table 2.1. For reference, the grey dashed line illustrates the true population average daily expenditures of 39.29789.

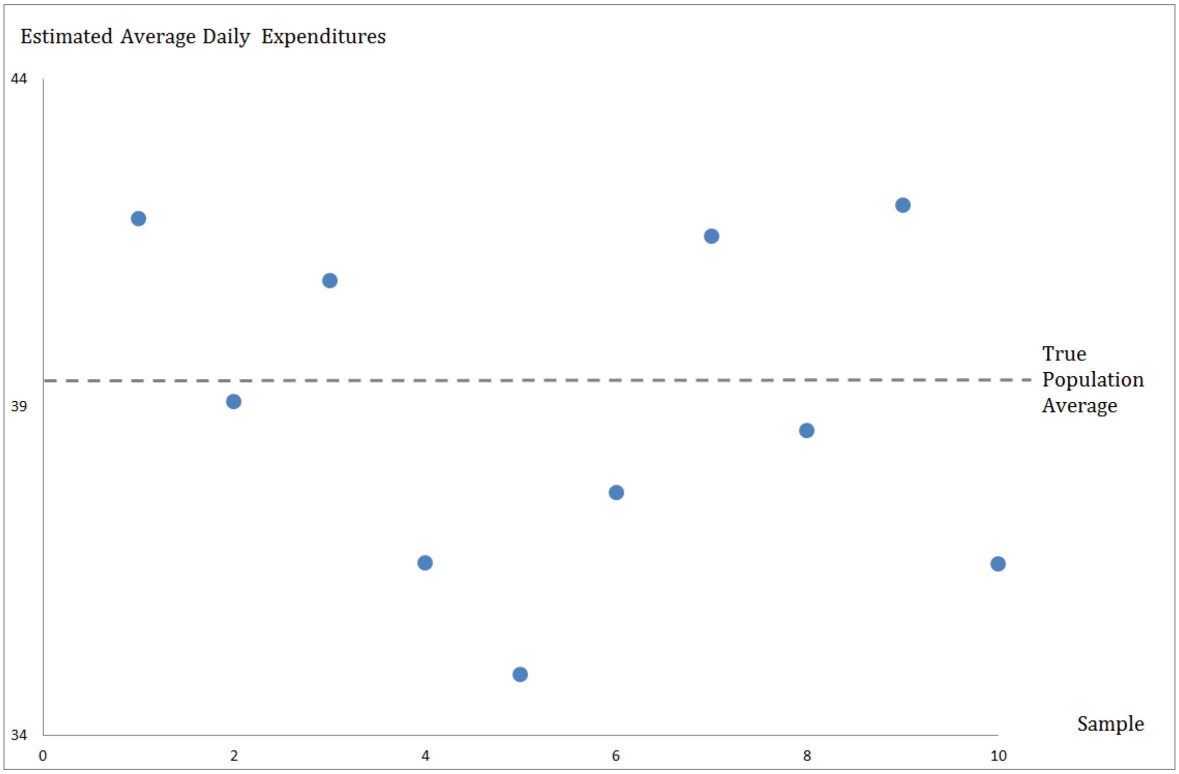


Figure 2.7. Estimates of average daily expenditures: Sample size 10 households

Table 2.1. 10 estimates out of 1,000 with a sample size of 10 households

Sample	Estimate
1	41.87220974
2	39.08382511
3	40.92349987
4	36.62444057
5	34.92545261
6	37.6951786
7	41.59581852
8	38.63977356
9	42.07090969
10	36.6068243

We have seen that the estimate of the population average daily expenditures will vary from sample to sample. Across the 1,000 samples of size 10, the standard deviation of the estimates was 3.2644308. The largest estimate across the 1,000 samples was 51.824298, while the smallest was 29.692244 — so there is quite a lot of sample by sample variation (called sampling variation) in the estimates. However, the average value for the estimates across the 1,0000 samples was 39.33052. This is very close to the true population average of 39.29789. Thus, while the estimates may vary in value from sample to sample, the central tendency of this stream of estimates is essentially the true population average. This is unbiasedness in action: The estimator may deliver a great deal of variation in estimate values from sample to sample, but it is "right on average."

Before moving on, we might consider a variation on this simulation exercise: What would happen if we change the sample size from 10 to 750? This is illustrated in Figure 2.8, which essentially repeats Figure 2.7 for the case of one example of 750 selections.

To understand the implications of increasing the sample size to 750 households, we draw another 1,000 samples, but in this case draw samples of size 750. The specific procedure is, unsurprisingly:

1. Draw a value from the uniform distribution for each of the 5,000 simulated households.

2. Select the 750 households with the smallest draws from the uniform distribution.

3. Record the average daily expenditures for that sample of 750 households.

This is thus exactly the same as the preceding exercise, just with a new target sample size of 750 instead of 10.

In Table 2.2, we present the estimates of population average daily expenditures from just 10 of those samples (randomly chosen from the 1,000). As we can see, once again the estimates still vary from sample to sample. Notice, however, that the estimates in Table 2.2 seem to vary in a narrower band than those in Table 2.1. This is visually illustrated in Figure 2.9, which essentially augments Figure 2.7 by adding the 10 observations from Table 2.1 (represented by the red dots).[16] Clearly, the estimates for the sample size of 750-households appear to vary far less from sample to sample than those associated with the sample size of 10 households. In the 750 household sample size case, the estimates of population average daily expenditures have a standard deviation across the estimates of .34883847, with a minimum estimate value across the 1,000 samples of 38.199 and a maximum value of 40.337785 — this is numerical evidence of the smaller sampling variation at the 750 household sample size. The average of the 1,000 estimates with a sample size of 750 households is 39.287134, which is once again extremely close to the population value of 39.29789.[17]

[16] In the interest of transparency, there is no relation between the sample numbers in Figures 2.7 and 2.9 and Tables 2.1 and 2.2. The 1,000 samples of size 10 were drawn from the 1,000 samples of size 750 in a completely arbitrary fashion and, in both cases, the 10 illustrative sample estimates were selected randomly and arbitrarily assigned a sample number. This does not in any sense alter the basic points that while the estimates associated with the samples of size 750 households also vary from sample to sample, they do so far less than those generated by sample sizes of 10 households.

[17] It is slightly closer to the true population value than the average across the averages generated from 1,000 samples of 10 households. This largely reflects the smaller standard deviation for the average of the averages in the 750-household sample size case. Since the sample by sample averages in the 750-household sample case vary less around the true population average, the average of them is a more precise estimator of the true population average of 39.29789. While unbiasedness holds regardless of sample size, fewer samples are typically required to make this clear when each involves a larger underlying sample size that generates a more precise estimate of the population average from each sample.

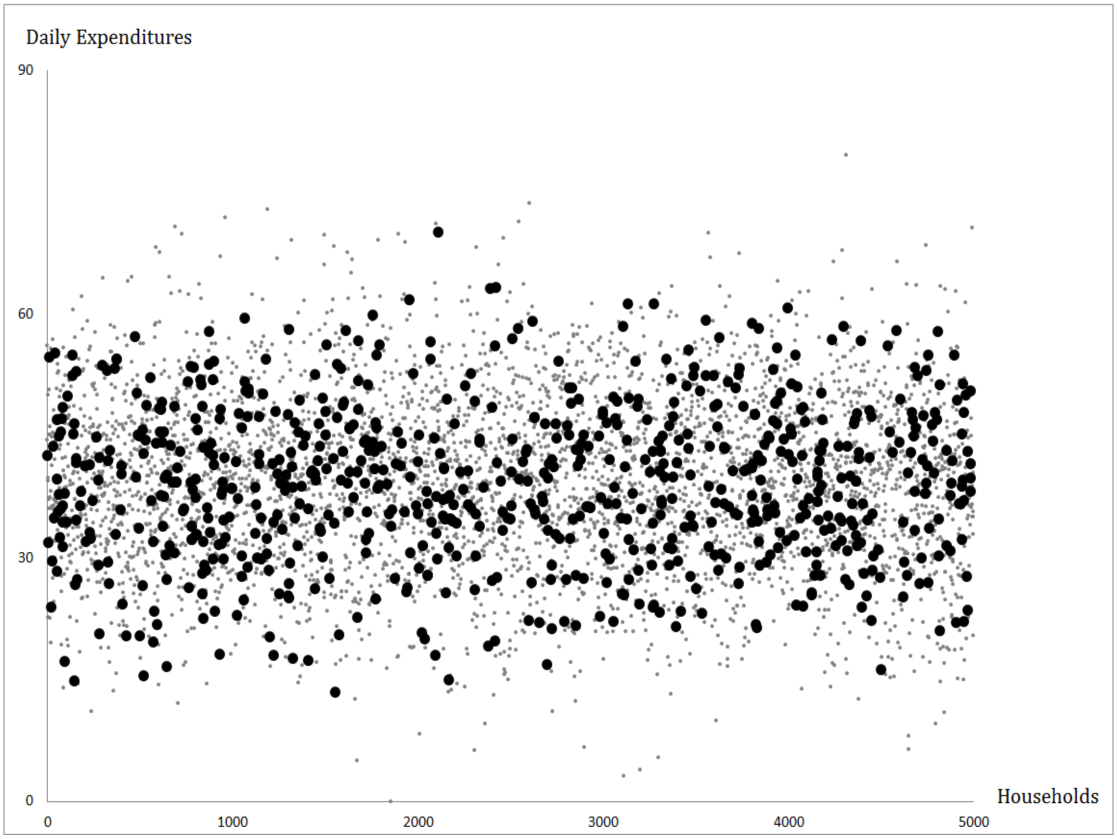


Figure 2.8. A sample of 750 from the population of 5,000

This illustrates once again the unbiasedness of the basic estimator and the fact that unbiasedness does not depend on sample size.

Table 2.2. 10 estimates out of 1,000 with a sample size of 750 households

Sample	Estimate
1	39.14759631
2	39.39117597
3	39.89114688
4	39.02860937
5	39.64354097
6	38.68658215
7	39.0396557
8	39.40299394
9	39.45032388
10	39.12667923

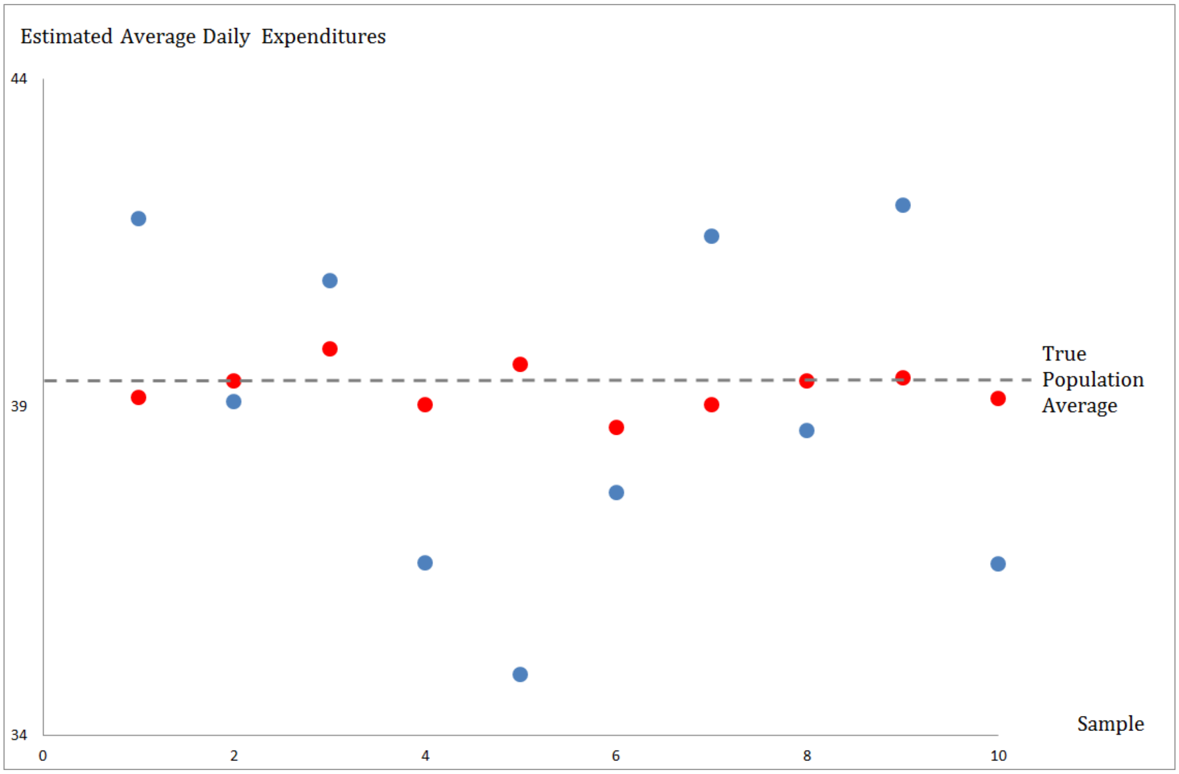


Figure 2.9. Estimates of average daily expenditures: Sample size 10 (blue dots) and 750 (red dots) households

We now bore in a bit more on the method of selection. In the preceding examples (of selecting samples of size 10 and 750 from the population of 5,000) the method of selection was random: Each member of the population was given a draw from the uniform distribution, and the households with the smallest 10 or 750 draws (depending on the target sample size) were selected. Because the draws from the uniform were random,[18] the chances that any household's draw was in the bottom 10 or 750 was also random. Selection was unrelated to any systematic characteristic of the household. Thus, households were effectively sorted into the sample in a fashion that had nothing to do with their characteristics: A wealthy (by daily expenditures) household was as likely to end up in the sample as a poor one.

The resulting samples on average reflected the distribution of different types of households in the population, and the simple mean of daily expenditures across the sample was an unbiased estimator of daily expenditures. For instance, for these simulations selecting 1,000 samples, we create three dummy variables for each member of the population that equal 1 if the household has a daily expenditure in the bottom 25, top 50, and top 25 percent of the distribution of such expenditures

[18] Actually, the draws were *pseudo-random*: Computer random number generators such as those employed for our numerical examples do not actually yield completely random numbers. If we think about a perfectly random number generator, one property that it should satisfy is that in a sequence of K draws $\{x_1, x_2, ..., x_K\}$ where x_i is the i^{th} draw, we should not be able to predict any draw in the sequence. Unfortunately, this is not exactly the case with most computer random number generators of which we are aware. Generally, random number generators rely on a seed. The initial value for that seed is s_1. Then, crudely (for those who want a more sophisticated discussion for the case of STATA, see `http://www.stata.com/statalist/archive/2012-10/msg01129.html`), the process of drawing a sequence of numbers works like this:

1. Using the 1^{st} seed value s_1, the first draw x_1 and seed value s_2 are generated.
2. Using the 2^{nd} seed value s_2, the second draw x_2 and seed value s_3 are generated.
3. ...
4. Using the $(K)^{th}$ seed value s_K, a new draw x_K and seed value s_{K+1} are generated.

The process is thus an iterative one across which depends on the initial seed and, for the second seed value s_2 onwards, generates new seed values. Let us see how this works with a simple STATA example (from Chapr2-2.do) where we set a seed value and then make three draws from the uniform distribution:

```
. set seed 12345
. di uniform()
.35762972
. di uniform()
.40044262
. di uniform()
.68938332
```

If we repeat this exercise, the results are:

```
. set seed 12345
. di uniform()
.35762972
. di uniform()
.40044262
. di uniform()
.68938332
```

In other words, re-setting the seed to its original value and redrawing from the uniform distribution three times yields the same sequence of values. The values in the sequence are thus determined by the first seed value (in this case 12345). The sequence of draws is thus not truly random but a deterministic function of the first seed value. Another curiosity of this is that eventually the generator will update the seed value to 12345 and the sequence will then repeat. This typically requires a vast number of draws but is yet again inconsistent with true randomness.

These details are by and large a curiosity. However, in this manual we will appeal frequently to simulations based on random number generation, and so these details are worth bearing in mind — the simulations are not exactly random.

across the population of 5,000 households (and 0 otherwise). The means of these three across the selection of 1,000 samples of size 10 households were 24.68, 50.37 and 24.96 percent, respectively, and 25.003067, 49.982667 and 24.971467 percent, respectively, for the exercise involving samples of size 750 households.

In other words, for both the 10- and 750-household sample size cases, the average shares of the samples of those from the bottom 25, top 50, and top 25 of the population-level distribution of daily expenditures is right in line with what they should be in terms of the samples reflecting the population. Thus, in terms of the only type household characteristic that we observe (so far), our samples appear on average to exhibit a general distribution of "types" that matches the actual share of these types in the population.

Because our samples reflected on average the appropriate distribution of types of households (from the standpoint of the actual share of those types in the population), the samples on average represented the population. Because our samples represented the population on average, so too did the estimates of daily expenditures emerging from them. As we have already learned, however, if the flow of estimates of the population parameter are "right on average," then they are unbiased.

Random selection was key to this good sampling performance. Because the samples were randomly selected and each household had some probability of selection, there was no type of household systematically excluded from the sample (which would have yielded samples that were distorted representations of the population from which they were drawn because there would have been certain types of households present in the population but not in the sample).

Moreover, under our rather simple method of selection (drawing a value from the same uniform distribution for each household and then choosing the households with the smallest draws), all households had the same *ex-ante* probability of selection. In the case of selecting samples of size 10, the *ex-ante* probability of selection for each household is 10/5,000=.002. For the sample size of 750 it was 750/5,000=.15. Samples in which all units have the same probability of selection are called "self-weighting" samples, and as we will see this is a fairly desirable property.

From an intuitive standpoint, the equal *ex-ante* probability of selection meant that no household had more likelihood of selection than any other. Because of this, the different types of household had no greater probability of selection than their shares in the population suggested they should. The resulting samples thus had (at least on average) the same distribution of household types as the population. In that sense the samples were representative of the population.

These examples (i.e., the 10- and 750-household sample size simulations discussed above) in some sense reflected the ideal of randomization: Random selection probabilities that were the same across all members of the population and unrelated to their characteristics. To see what the various properties of this simple sampling process meant for the results we have seen, we now discuss a series of variations on our simulation exercise that screw up one or more of these conditions. We can think of four possible (sometimes extreme) departures from the happy set of circumstances these simulations represented:

1. **Departure 1:** Wholly non-random selection

2. **Departure 2:** Selection that is random with the same probability of selection across types of sampling units with any positive probability of selection, with the possibility that it is zero for some types

3. **Departure 3:** Selection that is random and unrelated to the characteristics of the sampling units, but with selection probabilities that are nonetheless unequal across units of selection

4. **Departure 4:** Selection that is random but with probabilities correlated with the characteristics of the sampling units

We briefly consider each case with a simulation. Since the distinction between larger and smaller samples is not really very important for illustrating these cases (in the sense that in the discussion above we considered the 10- and 750-household sample size cases simply to illustrate the smaller sampling variation to the estimates generated by the latter), we will focus on a compromise sample size of 250 households for each of the simulations around these cases of potential complications.[19] However, we preserve the benchmark of considering performance across 1,000 simulated samples. In each example of departure our population is the same simulated 5,000 household population that has been the foundation of our discussion up to this point.

We begin with the case of wholly non-random selection (**Departure 1**). In the STATA file Chap2-3.do we consider this rather extreme case by selecting our samples of 250 households as follows:

- Sort the households in the population of 5,000 according to daily expenditures.
- Select the first 250 households per this ordering.
- Record the average daily expenditures for that sample of 250 households.

The results from repeating this process 1,000 times are stark. The average estimate across the 1,000 samples was 17.98697, far from the true population average daily expenditures of 39.29789. This is not a result of, for instance, wild sampling variation leading to an average estimate that is still not close to the true population value even across 1,000 samples: Increasing the number of replications to 100,000 yields an average of the estimates across the 100,000 samples of 250 households of 17.98697, which is in fact the same average estimate yielded by the 1,000 sample replication exercise. There is a reason for this: When one selects in this fashion the same 250 households will be selected *every time*. Unsurprisingly, 100 percent of this sample of 250 households is from households in the bottom 25 percent of all households in terms of the population distribution of daily expenditures, with none coming from households from the top 50 or 25 percent at the population level. Because the same households appear in each sample, the estimate of average daily expenditures is the same for each sample and hence the true sampling variance is essentially 0.[20] This sampling scheme has yielded

[19] The specific sample size used is not important, however.

[20] That said, conventional indicators of sampling variance such as those provided by STATA's `mean` command will fail to recognize this. For instance, in each sample STATA's `mean` command reports a standard error estimate for the estimate of population average daily expenditures of 0.2407259. This is a useful cautionary tale about naive

a completely distorted representation of the population, with all members of the sample coming from the lowest expenditure households (indeed, from the bottom 5 (=100*(250/5000)) percent of households) and none coming from the top 50 or 25 percent of the population distribution of daily expenditures.

This is a an extreme example in several ways. Nonetheless, it is useful in being crudely extreme. In this case, perhaps the most meaningful way of looking at the problem is that the households in the top 95 percent of the population distribution of daily expenditures had zero probability of selection. In general, unbiased estimation requires all elements of the population of interest to have a positive probability of selection into the sample.

The other sense in which this was an extreme example is that the household's value for the outcome perfectly predicted inclusion into the sample. In real world applications, we often run into a more intermediate case, whereby the outcome value is associated with the probability of selection into the sample but does not wholly determine whether a member of the population is selected into the sample. We will consider this intermediate case later, when we discuss Departure 4.

To sharpen our sense of the importance of all members of the population having a positive probability of selection into the sample, we turn to **Departure 2**: selection that is random with the same probability of selection across types of sampling units with any positive probability of selection, with the possibility that it is zero for some types. The simulations for this case are based on STATA code Chap2-4.do. Once again, our focus is the same population of 5,000 households that we have considered in the previous simulations. The true mean of daily expenditures across this population is thus still 39.29789. The process of sample selection proceeds as follows:

1. Draw a value from the uniform distribution for the 5,000 simulated households, except the 750 households with daily expenditures between the 75th and 90th percentile of the population distribution.

2. Select the 750 households with the draws from the uniform distribution with the smallest values.

The little twist here is that some households (most specifically those in the 75^{th} through 90^{th} percentiles of the population distribution for daily household expenditures) receive no draw from the uniform distribution. Since the draws are the basis for determining selection into the sample, in practice this means that these households have no chance of selection into the sample. Of the 4,250 households that do have a chance of selection into the sample, they all enjoy the same *ex-ante* probability of selection: 5.882353 (=100*(250/4,250)) percent.

This is a clear consequence to omitting some households. Repeating the sampling procedure for Departure 2 1,000 times, we find that the average estimate of average daily expenditures across

application of formulaic estimates of sampling variation — they assume a sampling process. In this case, the actual selection mechanism did not conform to the assumptions behind the formula for the standard error estimate. One should never take for granted that the sampling variation estimates provided by the default formulas in "canned" statistical packages are necessarily correct.

these 1,000 samples is 37.604046, against a true population value of 39.29789.[21] This bias (i.e., the difference between the average of the estimates of 37.604046 and the population value of 39.29789[22]) reflects the omission of a type in the population (those households with daily expenditures from the 75^{th} to 90^{th} percentile of the distribution of this outcome across the population) from the samples. This manifests itself in the average representation across samples of households from the bottom 25, top 50, and top 25 percent of the population distribution of daily expenditures at 29.374, 41.204 and 11.8492 percent, respectively. Thus, those in the bottom 25 percent are over-represented in the samples relative to their share of the population, while those in the top 50 and 25 percent are somewhat under-represented.

Interestingly, we can get the average estimate closer to the true population value by applying a weight to the observations that are selected into the sample. After some fiddling, the authors found that applying a weight based on the formula

$$w_i = \frac{1}{\sqrt{Y_i}} \cdot .05$$

(where Y_i is the daily expenditure of household i from the sample of $i = 1, ..., 250$ households) yielded an average estimate of the population average daily expenditures of 39.004847, which is indeed closer to 39.29789 than the average of the unweighted estimates of the average, which was 37.604046. Applying this weight was straightforward. We simply estimated population daily expenditures in each sample by

$$\frac{\sum_{i=1}^{250} w_i \cdot Y_i}{\sum_{i=1}^{250} w_i}$$

In other words, this formula replaces the equal weight of 1/250 for observation with the observation-specific weight w_i.

Unfortunately, this remedy would be impractical in real world practice as the authors came to this weight by a process trial and error that relied crucially on knowing the actual population average. Because they knew the population average, they could keep experimenting with new weight formulas to see whether they yielded an average of the estimates across the samples closer to the

[21] This situation does not substantially improve by increasing the number of samples to, for instance, 10,000. This yielded an average of estimates across the 10,000 samples of 37.587017.

[22] One could reasonably ask how we decide in the context of a simulation whether unbiased estimation is occurring since in no case does the average of the estimates *exactly* equal the population value of 39.29789. There is admittedly some hand waving to this, but the basic decision-making rule is that the sampling process yields unbiased estimates if the average of the estimates is clearly converging to the true population value as we increase the number of samples of the same size that we simulate. For instance, in our canonical original case of drawing 1,000 samples of size 10 the average of the estimates across the 1,000 samples was 39.33052 (which, to be fair, is much closer to the population value of 39.29789 than, for instance, the average of 37.604046 recovered from our Departure 2 simulations). However, if we decrease the number of simulations for the original example to 100, the new average across the estimate of the average for the 100 samples is 39.135647, which is a little further from 39.29789 than the 39.33052 based on 1,000 samples but still very close.

Technically speaking, another way to assess unbiasedness would be to repeat the simulation of 1,000 samples many times, and take the average of the average of the estimates across those many simulations of 1,000 samples.

In any case, an admitted defect to our rather casual approach to assessing unbiasedness in the context of our simulations is that it could admittedly confuse a *slightly* biased estimator with an unbiased one. We ask for the readers' patience with our hand waving around the possibility of slight bias.

true population average. The reason that experimentation was required is that classical sampling theory did not suggest a clearly preferred formula for the weights. As we will see later, classical theory does suggest a formulaic remedy when everyone in the population has some probability of selection into the sample.

Departure 2 is in some sense a refinement of Departure 1. The outcome of interest (daily expenditures) does not wholly determine which households are in the sample, but certain types of households cannot get into the sample.

Selection of samples is usually done from lists of members of a population called sampling frames. (Sampling frames will in short order be the subject of much discussion in this manual.) This is effectively an example of frame *under*-coverage: Some households are kept off of the list from which households are selected.

In real world applications, the problem of frame under-coverage is often more subtle. For instance, presented with the set of households from which selection of the samples was made in this example, most samplers and analysts would notice a glaring gap in the distribution of daily expenditures in the sample. More subtly, suppose that those in the 75^{th} through 90^{th} were simply under-represented in the frame compared with their actual numbers in the population. How would we be able to tell that this might be happening?

An example of this is provided in a slight variation on Departure 2 available in STATA code Chapter2-4a.do. This variation is essentially the same as Departure 2, except that rather than excluding from the samples all households in the 75^{th} to 90^{th} percentiles of the population distribution of daily expenditures, we exclude only half of those households (randomly selected).

Specifically, the simulation proceeds as follows. First, the same population of 5,000 households is simulated. Then a draw is made from the uniform distribution for each of the 750 households in the 75^{th} to 90^{th} percentile of the population distribution of daily expenditures. Of those 750 households, the half (375) with the smallest draws is excluded from the possibility of selection into the samples. We then repeat the following exercise 1,000 times:

1. Draw a value from the uniform distribution for the 4,645 simulated households still eligible for selection into the samples (4,250 households outside of the 75^{th} to 90^{th} population percentiles for daily expenditures plus the 375 households from the 750 households in those percentiles that were randomly assigned to be at risk of selection into the samples).

2. Select the 250 households with the draws from the uniform distribution with the smallest values.

Since the backdrop to these 1,000 sample selections is the same reference population of 5,000 households (regardless of who is eligible for selection into the samples), the true population average daily expenditures is 39.29789.

The average across the 1,000 samples of 250 households of the estimates of that population average at 38.515654 falls short of the true population value. However it does so less than the average of 37.604046 from the original Departure 2 (when all of the households from the 75^{th} to

90^{th} percentiles were excluded from selection into the samples). As in this original Departure 2 simulation, the reasons are similar — the average share of households from the 1,000 samples from the bottom 25^{th}, top 50^{th}, and top 25^{th} of the population distribution of daily expenditures is 27.06, 45.9484 and 19.0308 percent, respectively. In other words, as in the original Departure 2 case, the upper end of the population distribution of daily expenditures is under-represented in the samples (relative to their share in the population), though not as dramatically as in the original Departure 2 case.

Nonetheless, the estimates are not unbiased in this instance. The reason is that there are still members of the population with no probability of selection. Once again we fiddled around with possibilities for weights. When we apply the weight

$$\frac{1}{Y_i^{.25}} \cdot .05$$

to the calculation of the the estimate of the population average, the average of those estimates across 1,000 samples is 39.231514, which is closer to the population value of 39.29789. Thus, weights did get our estimates closer to the true population average. However, as in the original Departure 2 case, this would not be a feasible solution in practice since the weight formula was arrived at by trial and error, a process that was only productive because we knew the true population average and hence could judge whether different candidate weight formulas got us closer to that population average.

In situations akin to this (some subpopulation groups are effectively unrepresented in the frame from which samples are selected compared with their true share in the population) there is a remedy called post-stratification adjustment that can sometimes be pursued. Essentially (and admittedly crudely), post-stratification involves comparison of the distribution of sample characteristics to the true (or "reasonably true") population distribution, which is usually based on consultation of some auxiliary information source (such as a census or a survey based on a more complete frame), and adjustment (via weight modification) for any discrepancies with the population distribution of those characteristics. Post-stratification, a topic we will discuss more later in the chapter, is unsurprisingly also used as an adjustment for non-response out of those selected for the sample.

At this point we digress somewhat to consider another variation on Departure 2. This variation is contained in STATA code Chapter2-4b.do. This variation is essentially the same as Departure 2, except that rather than excluding from the samples all households in the 75^{th} to 90^{th} percentiles of the population distribution of daily expenditures, we exclude 750 households randomly selected *from across the entire population.*

Specifically, the simulation proceeds as follows. First, the same population of 5,000 households is simulated in the same fashion as in earlier examples. Then a draw is made from the uniform distribution for each of the 5,000 households in the population. The 750 households with the smallest draws are excluded from selection into the subsequently simulated samples. We then repeat the following exercise 1,000 times:

1. Draw a value from the uniform distribution for the 4,250 households assigned to be at risk of

selection into the samples.

2. Select the 250 households with the draws from the uniform distribution with the smallest values.

Once again, some members of the population are excluded from selection into the samples (in other words, they have zero probability of selection). However, this variation introduces the twist that exclusion is now entirely random: These 750 households are randomly selected for exclusion. Hence exclusion is unrelated to the characteristics of the households in the population (such as daily household expenditures).

Since the backdrop to these 1,000 sample selections is the same reference population of 5,000 households (regardless of who is eligible for selection into the samples) the true population average daily expenditures is 39.29789. The average of the estimates of average daily expenditures across the 1,000 simulated samples is 39.308808: From a numerical standpoint this sample selection procedures yields essentially unbiased estimates of the true population mean.

At first it might seem hard to understand this in light of the problems that exclusion of members of the population from selection causes in the earlier simulations for Departure 2. The key in this case is that the excluded households were excluded randomly. Thus, the set of households from which we selected the 1,000 samples of 250 households can be viewed in some sense as a randomly selected 4,250 household subsample of the population. Since it was randomly selected it was representative of the population. Thus, the 250 household subsamples selected from this 4,250 household sample were also representative of the population. Randomly selected subsamples of randomly selected samples are, in general, also representative of the same population for which that original sample was representative.

So this would seem an instance where excluding some households from selection is without practical consequence (at least in terms of the estimate of average daily expenditures). However, a strong note of caution must be sounded — we should not get comfortable with the idea of a sampling process that excludes members of the population from the possibility of selection into the sample. First, the entire "exclusion at random" explanation for arguing that there are no consequences to excluding population members is a heroic assumption for general practice and typically untestable. If members of the population are excluded from selection (perhaps because the list of population members from which selection is made is incomplete) and the sampling personnel for a survey are aware of this, they should not simply assume that the exclusion was random.

Second, in this instance we are estimating an average. For the purposes of estimating many other kinds of population parameters, any kind of exclusion of members of the population from eligibility for selection would be a problem. Consider, for instance, the challenge of estimating total daily expenditures for the population of households. Suppose that we selected a sample of 250 households by simple, equal probability of selection means (such as ranking households by their value for a draw from the uniform distribution and then choosing the 250 households with the smallest draws). The appropriate sample estimator of the population total would be

$$\frac{1}{p} \cdot \sum_{i=1}^{250} Y_i$$

where p is the percentage *of the population* selected. If all 5,000 households had been eligible for selection, this would have been .05 and the estimate of the population total would then be

$$\frac{1}{.05} \cdot \sum_{i=1}^{250} Y_i = 20 \cdot \sum_{i=1}^{250} Y_i$$

This simply says that the sample total must be multiplied by 20 to obtain the population total since the sample represented only one twentieth of the population. Suppose, however, that only 4,250 households out of the 5,000 in the population were actually eligible for selection and we did not know this. Our estimate would then be

$$\frac{1}{\frac{250}{750}} \cdot \sum_{i=1}^{250} Y_i \approx \frac{1}{.0588} \cdot \sum_{i=1}^{250} Y_i \approx 17 \cdot \sum_{i=1}^{250} Y_i$$

Our estimate of the population total will systematically underestimate the actual population total by a factor of 15 ($= 100 \cdot (1 - 17/20)$) percent. Assuming "exclusion at random" would not bail us out this time.

Thus far, we have seen an instance where the sample selection process fell short of the simple ideal of the original simulation example (involving selection of samples of size 10 and 750 households) by excluding some households in the population from selection into the samples. When this exclusion was related to the outcome (that is, population members were not excluded at random) the result was estimates that were biased (as evidenced by systematic departures of the average of estimates across repeated samples from the true population value). However, we were also able to compensate for this by crafting, via trial and error, weights.

When we say "trial and error," what we mean is that there really was no theoretically recommended or obvious formula for the weights. Rather, we just tried candidate functional forms for the weights until we finally found one that got us close to the true population value. For instance, we tried

$$\frac{1}{Y_i^a} \cdot .05$$

for values of α of .1,.2,.25,.3,.33, .5, 1, 2, 3 and 4. Sometimes the weight made things worse (i.e., yielded averages of estimates across samples even farther from the true population average than what we would have obtained by just ignoring weights altogether). Sometimes it made it better. As we have mentioned, this trial and error process required us to know the true population average and thus be able to assess whether things were getting better or worse with a candidate weight.

But what if (as will virtually always be the case when you conduct surveys) we don't know the true population value for a parameter? In this instance, trial and error would be useless and we would need to appeal to a theoretically valid formula. It turns out that such a formula exists. In instances where all members of the population have some probability of being selected and the i^{th} member of the population has a probability of selection of

$$p_i$$

the appropriate weight to apply to that member of the population (if selected) is

$$w_i = \frac{1}{p_i}$$

In other words, the weight for the i^{th} unit is the inverse of their probability of selection. The estimator for the population average once weights are applied is

$$\frac{\sum_{i=1}^{n} w_i \cdot Y_i}{\sum_{i=1}^{n} w_i}$$

where n is the sample size. The weights serve either to amplify or reduce the "information contribution" of each member of the sample depending on whether that members value respectively exceeds or falls short of the average weight across the sample.

This makes intuitive sense. We have suggested that an ideal sampling process would result in a self-weighted sample. Every member of the population has the same probability of selection, and hence their chance of representation in the sample, as well as the chance of representation for the "type" of member of the population that they are, is the same as their actual share of the population. This allows estimates from such samples to reflect on average the distribution of types in the population and hence the average of the estimates across the flow of estimates from repeated samples to be "right on average" or, in other words, unbiased.

However, if different members of the population have different probabilities of selection, then the weights proposed (whereby the weight is the inverse of the probability of selection) compensate for these imbalances. Those members with a higher probability of selection will be over-represented in samples (compared with their true share of the population). They are thus assigned a smaller weight for estimation, reducing their influence on estimates back to what it should be, given their share of the population. Those members with a lower probability of selection will be under-represented in samples (compared with their true share of the population). They are assigned a larger weight for estimation, inflating their influence on estimates back to what it should be given their share of the population.

Graphically, consider Figure 2.10. In Figure 2.10, we illustrate a population with two types: red types and blue types. From left to right, the vertical bars are the representation of the two types in the population, a sample from that population, and the information flows behind a calculation of an estimate of some parameter (such as the average, rate, or total) of something across the population). Beginning with the left-most bar, the red and blue types have equal shares (specifically, half each) in the population. However, selection for a sample from that population is then done in such a way that the blue type was twice as likely to be selected into the sample as the red type (the middle vertical bar). The problem now is that the raw sample provides a distorted picture of the population, with blue types being over-represented compared with their true share in the population. By the same token, red types are under-represented in the raw sample. Weights that would allow representative estimates to be formed from the sample would have to reduce this distortion to the raw sample in computing the estimate. They do this by reducing the contribution or information flow made by the blue type in the estimation process until it is in line with their true representation in the population. Similarly, they amplify the information contribution of red types. This is illustrated in the final vertical line.

Departures 1 and 2 focused on a somewhat different case: members of the population with no (zero) probability of selection into a sample. In the simulations for Departures 1 and 2, all members of the population with positive probability of selection had the same probability of selection (i.e., equal probability of selection). This allowed us to focus on the problem of some members of a population having no chance of selection into a sample. As we saw, the consequence of this can be serious — a stream of estimates of the population parameter across samples that on average did not equal the population parameter value. In other words, the estimates were biased. The only remedy (the "trial and error" weight formulas) that improved the situation was an ad hoc, impractical one that would not be workable in the real world. Nonetheless, these examples, however hokey and atheoretical, introduced a powerful possibility — that weights could bend estimates back toward the true population value.

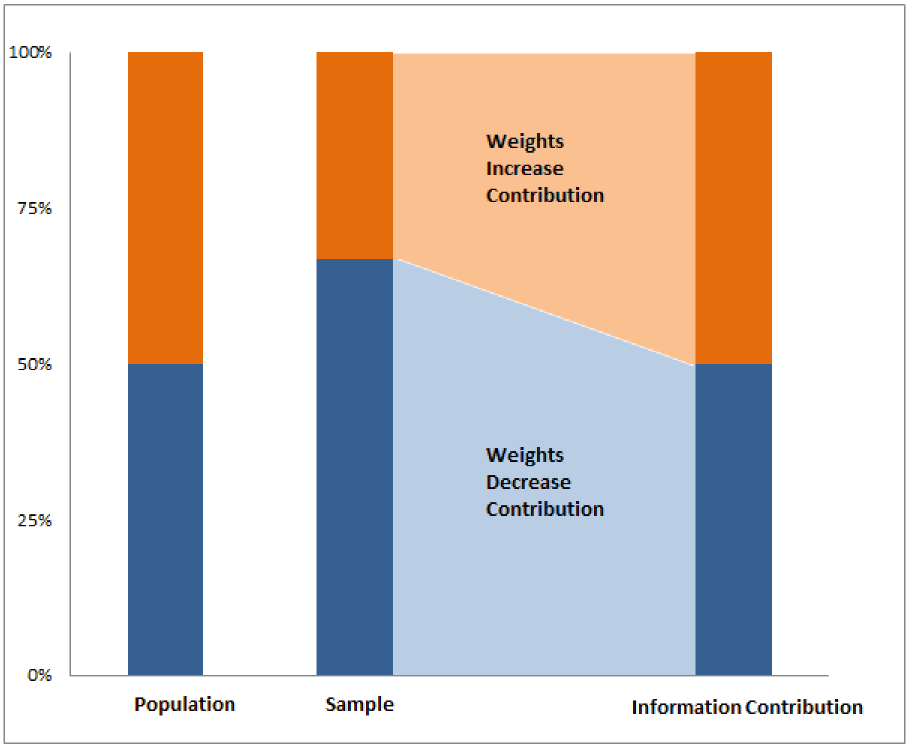


Figure 2.10. The effect of weights

Departures 3 and 4 essentially look at the consequences of uneven probabilities of selection and the use of weights to address them. To focus on this issue, we abstract away from Departures 1 and 2 by allowing all members of the population to have some positive probability of selection into a sample. However, the probabilities of selection are not the same across the members of the population. For the simulations that we perform around these departures, we continue to rely on the same simulated population of 5,000 households considered thus far.

Before proceeding to Departure 3, we need to introduce a different method of sample selection, called systematic random sampling. There are two reasons to do this. First, in real world practice sample selection is often (indeed typically) done by systematic random sampling. Second, as you will soon see, now that we are examining uneven selection probabilities, it will become necessary to model those selection probabilities in an analytically precise way. For the present purposes this is most readily achieved through systematic random sampling.

Before getting down to details, the basic idea of systematic random selection is very simple. Systematic random selection begins with a list of all of the sampling units in a population. For instance, in the current running example, this would mean a list of all 5,000 households in the population. We next determine something called the sampling interval. The sampling interval is the interval the separates selections on the frame. Call this I. We then choose a randomly selected sampling unit from the first I units on the list. This is the random start household, or RS. The next household we select is at position $RS + I$ on the list. Then we selected households

$$RS + 2 \cdot I, RS + 3 \cdot I, RS + 4 \cdot I, ...$$

until we have selected all of the households that we wish to select. In other words, the logic is that we first select a randomly chosen households, and then every I^{th} household after that.

Let's get a handle on this by working through an example. Consider our population of 5,000 households. The frame for this population is given in Table 2.3. As you can see, it is simply a list of all 5,000 households. It is typical practice to randomly sort the frame (for instance, by drawing a uniform random variable for each element of the frame and then sorting the sampling units by the size of that draw). The reason for this is to avoid any strange periodicity to the selections (as could possibly happen if the frame is an official frame and the order of units is not at all random but instead somehow grouped, as in by administrative units).

Table 2.3. The sampling frame

Household
1
2
3
4
5
6
...
4994
4995
4996
4997
4998
4999
5000

Suppose that we were to select 1,000 households from the population. (We choose this large number so that the sampling interval will be short, making it easier to demonstrate the selection process.) With 1,000 selections out of a population (or at least frame) with 5,000 sampling units, the sampling interval is

$$\frac{5,000}{1,000} = 5$$

The first step is then to choose a random start. Given the sampling interval of 5, this means selecting a number between 1 and 5. Suppose that we select 3. This means that the 3^{rd} household on the list is the first selected, as in Table 2.4.

Table 2.4. The first household selection

Household	Selected
1	No
2	No
3	Yes
4	
5	
6	
7	
8	
9	
10	
11	
12	
13	
14	
15	
16	
17	
18	
19	
20	
21	
...	
4996	
4997	
4998	
4999	
5000	

For our second solution, we add the sampling interval to the random start, $3 + 5 = 8$, to determine that the 8^{th} household on the list is the second selection. We then add the sampling interval to that to determine that the third selection will be the 13^{th} $(= 8+5 = 3+2 \cdot 5)$ household. By similar logic, the fourth selection will be the 18^{th} $(= 13 + 5 = 3 + 3 \cdot 5)$ household. This process of moving ahead from each selection by the margin of the sampling interval to determine the next selection continues until all 1,000 households have been selected. Based on the starting value of 3, we can easily calculate that the $1{,}000^{th}$ household selected will be the $4{,}998^{th}$ $(= 3 + 999 \cdot 5)$ household. These selections are illustrated in Table 2.5. Essentially, the only truly random part of this selection process is the random start. Every other selection is a deterministic function of the random start (RS) and the sampling interval (I) such that the k^{th} household selected is household number

$$RS + (k-1) \cdot I$$

Nonetheless, the random start alone is sufficient to guarantee that this is a random selection process.

Table 2.5. The full household selection

Household	**Selected**
1	No
2	No
3	Yes
4	No
5	No
6	No
7	No
8	Yes
9	No
10	No
11	No
12	No
13	Yes
14	No
15	No
16	No
17	No
18	Yes
19	No
20	No
21	No
...	...
4996	No
4997	No
4998	Yes
4999	No
5000	No

This systematic random selection exercise has been an example of equal probability selection, with each household having an *ex-ante* probability of selection of

$$\frac{1,000}{5,000} = .2$$

This type of selection is often referred to as equal probability or “epsem” selection.

In many cases, systematic random sampling involves unequal *ex-ante* probabilities of selection across sampling units. Typically, in these cases the selection probabilities for sampling units involve some sort of “size measure.” This is often the population size of the sampling unit, per some measure of population size. For instance, suppose that we could observe the household population (for instance, number of regular or “de-jure” members) for each household. A hypothetical frame containing this information is presented in Table 2.6.

Table 2.6. The sampling frame, with a size measure

Household	**Population**
1	7
2	5
3	13
4	8
5	11
6	8
...	...
4994	2
4995	1
4996	9
4997	13
4998	12
4999	5
5000	7

The basic logic of selection is not dissimilar to the equal probability of selection case, except that now the focus is on household size rather than sampling unit number on the frame. To begin with, assume that the total population of these households (or, in different terms, the total “size” of the frame) is 35,000 households. Suppose as well that we again wish to select a sample size of 1,000 households. The sampling interval is now the total frame size divided by the number of selections we wish to make

$$\frac{35,000}{1,000} = 35$$

Next, we must calculate the running cumulative size of the frame for each successive sampling unit, as in Table 2.7.

Table 2.7. The sampling frame, with a size measure

Household	Population	Cumulative Population/Size
1	7	7
2	5	12
3	13	25
4	8	33
5	11	44
6	8	52
7	7	59
8	10	69
9	9	78
10	12	90
11	10	100
12	11	111
13	9	120
14	8	128
...	...	...
4994	2	34,953
4995	1	34,954
4996	9	34,963
4997	13	34,976
4998	12	34,988
4999	5	34,993
5000	7	35,000

The next step is to choose a random start. As in the equal probability of selection case, the random start will be somewhere between 1 and the sampling interval (of 35). Suppose that we randomly select 12 as the starting point. This means that the observation for which the cumulative population contains this random start value is selected. In this case that is the second observation, since the start point of 12 is greater than the upper limit of the cumulative population figure of 7 for the first observation and is less than the lower end of 13 for the cumulative population figure for the third observation. This is illustrated in Table 2.8.

Table 2.8. The sampling frame, with a size measure

Household	Population	Cumulative Population/Size	Selected
1	7	7	No
2	5	12	Yes
3	13	25	
4	8	33	
5	11	44	
6	8	52	
7	7	59	
8	10	69	
9	9	78	
10	12	90	
11	10	100	
12	11	111	
13	9	120	
14	8	128	
...	...	...	
4994	2	34,953	
4995	1	34,954	
4996	9	34,963	
4997	13	34,976	
4998	12	34,988	
4999	5	34,993	
5000	7	35,000	

Following a logic similar to the equal probability of selection case, the second selection is the household for which the cumulative population figure contains

$$RS + I = 12 + 35 = 47$$

This is the 6^{th} household on the frame. The next (third) selection is then

$$RS + 2 \cdot I = 12 + 2 \cdot 35 = 82$$

It is the 10^{th} household on the list for which the cumulative population contains 82, and hence the 10^{th} household is the third selected. The fourth household selected is that for which the cumulative population contains

$$RS + 3 \cdot I = 12 + 3 \cdot 35 = 117$$

This is the 13^{th} household on the list. Selection thus continues on the rule that the k^{th} household selected is that for which the cumulative population contains

$$RS + (k - 1) \cdot I$$

The 1,000th household chosen is household the household for which the cumulative population contains

$$RS + 999 \cdot I = 12 + 999 \cdot 35 = 34,977$$

This is the 4,998th household. These selections are presented in Table 2.9. This type of selection is typically referred to as "probability proportional to size" (pps) selection or, more simply, "size sampling."

Table 2.9. The sampling frame, with a size measure

Household	Population	Cumulative Population/Size	Selected
1	7	7	No
2	5	12	Yes
3	13	25	No
4	8	33	No
5	11	44	No
6	8	52	Yes
7	7	59	No
8	10	69	No
9	9	78	No
10	12	90	Yes
11	10	100	No
12	11	111	No
13	9	120	Yes
14	8	128	No
...	...	...	
4994	2	34,953	No
4995	1	34,954	No
4996	9	34,963	No
4997	13	34,976	No
4998	12	34,988	Yes
4999	5	34,993	No
5000	7	35,000	No

One thing that is somewhat less obvious in the case of selection by size is the probability of selection for a given sampling unit. To fix ideas, suppose that the population of the i^{th} household is POP_i, while the total population across all households in the frame is POP. If we were to select just one household from the frame, we would do so by selecting randomly a number between 1 and

POP since, in the case of selecting one household, the sampling interval is of length

$$\frac{POP}{1} = POP$$

The probability that the i^{th} household is selected is then the probability that the random start (which in this case of a single selection also happens to be the "random stop") would be in the cumulative probability for this household. However, since that household's increment to the cumulative population is POP_i, this is

$$\frac{POP_i}{POP}$$

Now, if we were to select two households, this particular households probability of selection would be twice as great since it is the probability that either

$$RS$$

or

$$RS + \frac{POP}{2}$$

is in the cumulative increment to the total population associated with household i. The selection probability for the i^{th} household would then be

$$\frac{2 \cdot POP_i}{POP}$$

More generally, the probability of selection for the i^{th} household would be

$$\frac{k \cdot POP_i}{POP}$$

if we were to make k selections.

Having introduced systematic sampling, we now turn to **Departure 3**: selection that is random and unrelated to the characteristics of the sampling units, but with selection probabilities that are nonetheless uneven across units of selection. Our simulations for this departure are based on STATA code Chap2-5.do. To begin with, we simulate our reference population of 5,000 households in the same fashion as in the earlier examples, including their original expenditure value (hence the true "population" mean is still 39.29789 across the population of 5,000 households). We then simulate a population level for each household. To do so, we first draw a variable, $inc2$, for each household in exactly the same fashion as the expenditure variable and then create another variable v that is $inc2$ plus 9 times a draw from the standard normal distribution, $v = inc2 + 9 * \mu$ where μ is the draw for each household from the standard normal distribution. The reason for drawing $inc2$ in the fashion that we do will become a little clearer in the course of the discussion of Departure 4.

Finally, we create a new variable *pop* that is the 20 quantiles of v.[23] The result is a variable *pop*

[23] In other words, we sort v in ascending order, assign each value for v to equal sized bins, and then create a new variable *pop* such that smallest 250 values for v are in bin number 1, the next smallest 250 values for v are in bin

that runs from $1, ..., 20$ and represents household "population." Because this variable is rooted in $inc2$, and not the original household expenditure variable, it is unrelated to that original household expenditure variable.

Having generated the population of 5,000 household with their expenditure and population values, we then repeat the following sample selection simulation 1,000 times:

1. Select 250 households by equal probability of selection (epsem) using systematic sampling and estimate population daily expenditures across the sample by simple average of daily expenditures for the selected households.

2. Select 250 households by pps systematic sampling, with the household population as the size measure.

3. Compute the simple unweighted average for daily expenditures across this new sample of 250 households.

4. Calculate the *ex-ante* selection probability for each selected household i as

$$p_i = 250 \cdot (pop_i/POP)$$

   where pop_i is the population of household i and POP is the total population size of the frame (i.e., $\sum_{j=1}^{5000} pop_j$).

5. Calculate the sampling weight for each selected households as

$$w_i = \frac{1}{p_i}$$

   Higher probability households thus have a lower weight.

6. Using this weight, calculate the weighted average

$$\frac{\sum_{i=1}^{250} w_i \cdot Y_i}{\sum_{i=1}^{250} w_i}$$

   for each of the $i = 1, ..., 250$ households selected for inclusion into the new sample.

For each sampling replication, we collect the results and will discuss below what we found across these samples.

This is quite a sampling simulation so it is worth "parsing." In step 1, we select via systematic random sampling, but with equal probability of selection. This is to provide a set of reference results from the ideal case of equal probability selection against which we can assess the consequences of unequal selection probabilities and the application of weights based on those probabilities.

number 2, etc. For each household, *pop* is then equal to the bin number for that households draw for v.

Next (step 2), we select 250 households systematically by pps, with household population as the size measure. Because this is size sampling (in other words, based on a size measure that varies across households), this is an unequal probability of selection approach to obtaining a sample.

We then (step 3) calculate a straightforward (i.e., unweighted) average for these newly selected 250 households

$$\frac{\sum_{i=1}^{250} Y_i}{250}$$

This provides us with a chance to examine the consequences of ignoring weights when selection probabilities are unequal.

We then calculate the probability of selection as

$$p_i = 250 \cdot \frac{pop_i}{POP}$$

with pop_i and POP defined as in the bullet points. This is an *ex-ante* probability of selection, and it reflects the persistent advantage or disadvantage the household has in terms of their probability of selection owing to their population size, with larger households having a larger probability of selection.

Given this, we then applied the basic formula for generating weights from selection probabilities (usually called probability weights):

$$w_i = \frac{1}{p_i}$$

We have already discussed the intuition behind this weight formula. At this point we emphasize the theoretical (as opposed to ad hoc trial and error) nature of this formula. We did not continuously rerun the simulations with different formulas until we found this one. Rather, we simply programmed this formula, which is based on statistical theory, into the simulation. But this also means that applying this weight did not depend on knowing the true value of the population parameter beforehand!

Finally, we compute the weighted average across the selected 250 households:

$$\frac{\sum_{i=1}^{250} w_i \cdot Y_i}{\sum_{i=1}^{250} w_i}$$

This is a departure, but not a major one, from the simple sample mean.

The results of this are pretty interesting. First, we consider the average of the estimates (of average expenditures) across the 1,000 samples for the cases of equal probability of selection, unequal probability of selection but no weights applied to estimation of the population average from the sample, and unequal probability of selection with probability weights applied to estimation of the population average from the sample. These averages were 39.301138, 39.326585, and 39.342158, respectively. In other words, all three are essentially equal to the true population average of 39.29789 across the 5,000 households.

This is perhaps not surprising for the first and last set of estimates. We have already seen that simple, equal probability of selection sampling yields unbiased estimates. And we have mentioned that the whole point of weights is to bend (at the point of estimation) a sample not necessarily representative of the population back toward a (weighted) contribution to the estimate that reflects each selected observation's true share of the population. Hence, the finding that the average of the estimates from weighted estimation is essentially equal to the true population average is, in fact, theoretically reassuring.

However, if the purpose of weights is to bend a sample unrepresentative of the population back toward representativeness in the face of unequal selection probabilities, how did we also apparently get away with ignoring weights when estimating the population average with samples selected with the same unequal probabilities of selection? How did we do essentially as well without weights as with them in the unequal selection probability case? Or, put differently, why didn't weights make much of a difference?

The answer is that weights made little difference to the average of the estimates across samples because they were uncorrelated with the outcome under consideration, daily expenditures. Recall that the probability of selection was rooted in variables, $inc2$ and μ, that were unrelated to household expenditures. For instance, the correlation between $inc2$ and y_i across the population of 5,000 households in our simulation exercise was just 0.0065, while that between μ and y_i was just 0.001. The result is that, across the population of 5,000 households, the probability weight

$$w_i = \frac{1}{p_i}$$

(where $p_i = 250 \cdot pop_i / POP$, pop_i is the population of household i and POP is the total population of all 5,000 households) is essentially uncorrelated with household expenditures (the correlation is just -0.0141). Weights uncorrelated with the variable under analysis are called "uninformative" weights because they do not inform (or alter on average) the estimate. (By contrast, weights correlated with the variable under analysis are "informative" because they do influence the estimate.)

The intuition behind this is surprisingly simple. If weights are uncorrelated with the outcome or characteristic under consideration, then applying weights does not amplify the contribution of one particular type of selected observation (where type is defined in terms of their value for the outcome or characteristic under consideration). Therefore weights really aren't bending the sample in terms of making some *types* of selected observations more important than others in terms of what we really care about: the value of the outcome or characteristic under consideration. Weights had no impact because the sample was already representative of the population in the sense that the average distribution of the outcome or characteristic under consideration was the same as in the population.

That is not to suggest that weights have no consequence. Let us now consider the *variance* of the estimates across the 1,000 samples for the cases of equal probability of selection, unequal probability of selection but no weights applied to estimation of the population average from the sample, and unequal probability of selection with probability weights applied to estimation of the population average from the sample. These were, respectively, .39992835, .35071568, and .69797903.

For the purpose of thinking about performance, it is in some sense only ever meaningful to talk about variances in relative terms (that is, the variance of this estimator compared with the variance of that estimator). To fix ideas, we thus use the variance of the estimates generated by the simple equal probability of selection samples (which we define as $V^{EPSEM} = .39992835$) as the reference. We compare V^{EPSEM} to the variance across the samples for the estimates of average expenditures yielded by pps sampling but no application of probability weights to the estimation of average expenditures ($V^{PPS,NW} = .35071568$) and the variance across samples for the estimates of average expenditures yielded by pps sampling with application of probability weights to the estimation of average expenditures ($V^{PPS,W} = .69797903$. In particular, we focus on the ratios

$$\frac{V^{PPS,NW}}{V^{EPSEM}} = \frac{.35071568}{.39992835} = .87694628$$

and

$$\frac{V^{PPS,W}}{V^{EPSEM}} = \frac{.69797903}{.39992835} = 1.7452602$$

Clearly, the mechanism for sample selection and, in the event of pps sampling, choice to use weights is not without consequence in terms of the sampling variation of the estimates of average expenditures. The sampling variance actually improved by *not* applying weight in the unequal selection probability case but worsened substantially by applying them. What could be going on here?

Let's begin by thinking a bit about the simple unbiased estimator of the population mean. It is:

$$\frac{\sum_{i=1}^{n} Y_i}{n}$$

We have already discussed conditions under which this is an unbiased estimator. However, it is worth reiterating. The expectation of this estimator is

$$E\left(\frac{\sum_{i=1}^{n} Y_i}{n}\right) = \frac{\sum_{i=1}^{n} E\left(Y_i\right)}{n}$$

where n is the more general sample size. The final step exploits the aforementioned fact (in the earlier footnote discussing the properties of the expectations operator) that, for two random variables X_1 and X_2 and two constants c_1 and c_2,

$$E\left(c_1 \cdot X_1 + c_2 \cdot X_2\right) = c_1 \cdot E\left(X_1\right) + c_2 \cdot E\left(X_2\right)$$

At this point in the earlier proof, we assumed (and do as well now) that the expectation of y is not evolving across the n selections (in other words, $E(y_i)$ is the same for all $i = 1, ..., n$). This simply means that, however we select, we are not selecting in such a fashion that we are by our design changing the type of household selected with each successive selection. This is not the same as assuming that the households we select have the same characteristics as the overall population (for instance, we could consistently select only households with more than 4 members). This simply means that $E(y_i)$ equals some constant ω for all selections $i = 1, 2, ...n$. Then, we have

$$\frac{\sum_{i=1}^{n} E\left(Y_i\right)}{n} = \frac{\sum_{i=1}^{n} \omega}{n} = \frac{n \cdot \omega}{n} = \omega$$

Perhaps the key idea from the earlier discussion of unbiasedness for present purposes is that although the daily expenditures y_i for each household i has its own unique value, all emerge from the same population and hence have the same expectation: the true population parameter.

Invoking this basic logic again once more, we now turn to considering the variance of this estimator, and in the interest of directness and clarity do so in a rather straightforward fashion. We have

$$var\left(\frac{\sum_{i=1}^{n} Y_i}{n}\right) = \frac{1}{n^2} \cdot \sum_{i=1}^{n} var\left(Y_i\right)$$

where the last step exploits a key property of variances: for any two *independent* (and we assume that our selections are independent of one another[24] and hence there is no reason to believe that there is a relationship between the y_is in the sample) random variables X_1 and X_2 and any two constants c_1 and c_2, we have[25]

$$var\left(c_1 \cdot X_1 + c_2 \cdot X_2\right) = c_1^2 \cdot var\left(X_1\right) + c_2^2 \cdot var\left(X_2\right)$$

However, we assume that the variance of i is also not related to i in the same sense that the expectation wasn't. Rather, all of the households emerge from the population distribution and thus the variance is the same for all household's observations. We then have

$$\frac{1}{n^2} \cdot \sum_{i=1}^{n} var\left(Y_i\right) = \frac{1}{n^2} \cdot n \cdot var\left(Y\right) = \frac{var\left(Y\right)}{n}$$

This is basically the standard error for the estimate of the population mean.

Having established this baseline standard error for the estimate of the population mean emerging from simple independent random sampling with equal probability of selection, let us now consider estimation of the population mean with weights. The estimator of the population mean in this case is

$$\frac{\sum_{i=1}^{n} w_i \cdot Y_i}{\sum_{i=1}^{n} w_i}$$

From this we can derive a "formulaic" estimate of the impact of weights in sampling variance. The

[24] This means that the selection of one household did not depend on the selection of another.

[25] The more general result is that

$$var\left(c_1 \cdot X_1 + c_2 \cdot X_2\right) = c_1^2 \cdot var\left(X_1\right) + c_2^2 \cdot var\left(X_2\right) + 2 \cdot c_1 \cdot c_2 \cdot cov\left(X_1, X_2\right)$$

or even more generally

$$var\left(\sum_{j=1}^{J} c_j \cdot X_j\right) = \sum_{j=1}^{J} c_j^2 \cdot var\left(X_j\right) + 2 \cdot \sum_{1 \leq j < k \leq J} c_j \cdot c_k \cdot cov\left(X_j, X_k\right)$$

The independence of the Xs would imply that all covariance terms equal zero: $cov(X_j, X_k) = 0 \quad \forall j \neq k$.

variance of the weight average of daily expenditures is given by

$$var\left(\frac{\sum_{i=1}^{n} w_i \cdot Y_i}{\sum_{i=1}^{n} w_i}\right) = \frac{\sum_{i=1}^{n} w_i^2 \cdot var\left(Y_i\right)}{\left(\sum_{i=1}^{n} w_i\right)^2} = var\left(Y\right) \cdot \frac{\sum_{i=1}^{n} w_i^2}{\left(\sum_{i=1}^{n} w_i\right)^2} = \frac{var\left(Y\right)}{n} \cdot \frac{n \cdot \sum_{i=1}^{n} w_i^2}{\left(\sum_{i=1}^{n} w_i\right)^2}$$

In other words, what we have in this case is the variance from simple independent random sampling with equal probability of selection $(var(Y)/n)$ multiplied by the weight driven factor

$$\frac{n \cdot \sum_{i=1}^{n} w_i^2}{\left(\sum_{i=1}^{n} w_i\right)^2}$$

This term is greater than or equal to 1 (it equals 1 in any case where the weights are all equal and in particular in the case where $w_1 = w_2 = ... = w_n = 1$).[26] This is theoretical evidence that the variance using weights should exceed the variance not using them.[27]

An interesting twist on this is to derive the "effective" sample size in the face of weight variation. By "effective" in this case we mean the number of observations with no weights applied that would yield the same amount of information as the sample size n with weights applied. To begin with, and using the original notation of Kish, we have

$$1 + L = \frac{n \cdot \sum_{i=1}^{n} w_i^2}{\left(\sum_{i=1}^{n} w_i\right)^2}$$

where "L" stands for "Loss" (as in, for example, "loss of efficiency." The effective sample size is then

$$\frac{n}{1+L} = \frac{n}{\frac{n \cdot \sum_{i=1}^{n} w_i^2}{\left(\sum_{i=1}^{n} w_i\right)^2}} = \frac{\left(\sum_{i=1}^{n} w_i\right)^2}{\sum_{i=1}^{n} w_i^2}$$

However, since $1 + L \geq 1$, we know that

$$\frac{n}{1+L} \leq n$$

So the variation from the weights means that our effective sample size is less than our actual sample size.

Unequal weights generally elevate sampling variance.[28] The basic "Kish inspired" formula provides us with a sense of the likely general upper bound to the sampling variance consequences of weights when the weights are not correlated with the outcome.

Unequal weights are the product of sampling with unequal probabilities of selection, a common feature of survey design. By survey "design" we typically mean the selection process of a particular survey. The consequences of the "design" for sampling variance are usually summarized in a statis-

[26] That it is greater than or equal to 1 is actually a special case of a special case of the Cauchy-Schwarz inequality.

[27] This has been a sort of rough and ready poor man's derivation of this result. For further and far more elegant treatment see, for instance, Kish (1992).

[28] We say generally because it is typically true, but there are some cases where the reverse is true, which we touch on below.

tic called the Design Effect. Most broadly defined, design effects capture the sampling variance associated with two different survey designs. The design effect $deff$ is[29]

$$deff = \frac{V^{Design\ 2}}{V^{Design\ 1}}$$

where $Design$ 1 and $Design$ 2 are alternative designs or methods of selection and estimation for survey samples. If the design effect is greater than 1, then $Design$ 2 is said to be *more efficient* (i.e., have less sampling variance) than design $Design$ 1.

In formulating the design effect, the denominator in some sense represents the sampling variance under some reference design, relative to which we wish to understand the sampling variance implications of some alternative design. The reference design is usually simple independent sampling with equal probability of selection. The design effect is thus conventionally formulated as

$$deff = \frac{V^{design}}{V^{SIRSEPS}}$$

where

$$V^{design}$$

is the design variance or variance under the selection design and

$$V^{SIRSEPS}$$

is the design under simple independent random sampling with equal probability of selection (or SIRSEPS). The design effect thus formulated essentially summarizes the consequences of moving from SIRSEPS to whatever selection design other than SISREPS is pursued. The more the design effect exceeds 1 the more relatively inefficient is the design utilized. That said, the design effect can be less than 1 (in which case the design pursued actually yields efficiency gains over SIRSEPS).

In our discussion thus far, the most logical analog to this specific and common approach to the design effect would be

$$\frac{V^{PPS,W}}{V^{EPSEM}} = \frac{.69797903}{.39992835} = 1.7452602$$

In other words, there is a dramatic loss of efficiency from sample selection with unequal selection probabilities and estimation of average expenditures with weights based on the selection probabilities compared with selection with equal probability across sampling units.

That said, we have also introduced another ratio of sampling variances that could be described as a design effect:

$$\frac{V^{PPS,NW}}{V^{EPSEM}} = \frac{.35071568}{.39992835} = .87694628$$

This simply captures the relative efficiency of selecting with unequal probabilities of selection but then eschewing weights for estimation compared with selecting with equal probabilities of

[29] The square root of the design effect, usually called the *design factor* or $deft$, captures the relative standard errors of two alternative designs.

selection (epsem). Perhaps a little surprisingly, this design effect suggests that pps sampling with no consideration of weights in estimation can actually be a relatively efficient strategy (more on this below).

These design effects were calculated based on the actual sampling variance observed over the 1,000 simulated sample selections. In actual empirical practice (that is, selecting a single, actual real world sample and then estimating some population level parameter from it), we would not be able to do this. We would only have one sample and so could not really estimate the design effect in this fashion (in other words, by actually computing the variance in estimates across the samples under alternative designs).

In that case an estimate based on one sample must be used. Most commercial statistical packages like STATA offer some commands or tools for doing this. In STATA, the most straightforward way is the `estat effects` command. The average across the 1,000 simulated samples of the estimates of the design effect via `estat effects` is 1.8051767 (very close to the 1,000 sample estimate of 1.7452602). A few other estimators of the design effect have been proposed for the case where the "design" mainly involves the application of weights to address unequal selection probabilities. Both are (to our knowledge) the achievement of the famous sampling statistician Leslie Kish.[30] First, we have the estimator

$$deff = \frac{n \cdot \sum_{i=1}^{n} w_i^2}{\left(\sum_{i=1}^{n} w_i\right)^2}$$

derived earlier. This had an average value of 1.8433699 across the 1,000 sample selection simulations. Another that Kish developed is

$$deff = 1 + cv^2$$

where cv is the ratio of the variance to the square of the mean of the relative weights. The average of this measure across the 1,000 samples is 1.8467569.

While all three suggest a design effect in the near neighborhood of the 1,000 sample estimate of the design effect, they all exceed it somewhat and in any case differ from each other. Slight though the discrepancies may be, they present a good opening to make an important point: Design effect estimation is a topic in flux and even with current "estimation technology" involves a lot of subjective choices. One should be conservative in their interpretation of design effect estimates. In some instances it is prudent to assume that they might be an underestimate of the true figure.[31]

Whatever the true design effect (and the consensus would seem to suggest that it is likely greater than, say, 1.6) there is obviously a huge price to pay in terms of sampling variation for using weights. It should now have become somewhat clearer why equal probability selection resulting in a self-weighting sample is generally preferred — use of unequal weights in estimation drives up sampling variance.

For our penultimate reflection on Departure 3, we tie together several threads of the discussion

[30] See, for instance, Kish (1992) mentioned in an earlier footnote.

[31] One other circumstances where conventional design effect estimates should be viewed with caution is when they are estimated with small sample sizes. The authors have instances of implausibly large or small design effect estimates generated by small samples.

to this point in the chapter by considering weights applied in the epsem selection case. For each simulation, we select samples of 250 households from a population of 5,000 of them. The probability of a household being selected in any given selection simulation is this

$$p = \frac{250}{5000} = .05$$

The weight is the inverse of this probability of selection, or 20. This weight will be the same across all households.

We now attempt to apply this weight to the epsem selection to see what will happen. To fix ideas, we consider the 1,000th sample selection simulation. In the STATA code, in each simulated selection (and hence in the 1,000th one) the epsem selection estimate of population average expenditures for the sample was based on the formula

$$\frac{\sum_{i=1}^{250} Y_i}{250}$$

where i indexes the 250 households selected. For the 1,000th sample this yielded an estimate of population average expenditures of 39.42705. For the 1,000th sample we try as well the weighted estimator

$$\frac{\sum_{i=1}^{250} w_i \cdot Y_i}{\sum_{i=1}^{250} w_i}$$

using the weight $w_i = 20$.

Applying the weighted estimator to the 1000th sample yielded an estimate of population average expenditures of ... 39.42705. In other words, the two estimators yielded the same result and the application of weights made no difference whatsoever in this case. This is because the weight was the same for each household selected. As you complete your reading of Departures 3 and 4, think of it this way: The weights that emerge from epsem selection are in some ways the ultimate example of uninformative weights because they are not at all associated with expenditures. (And, indeed, the correlation between income and the weight is not even meaningful in this case since the weight does not vary).

This is not to say that weights can always be ignored in the epsem selection case. For one thing, they actually do have a behavioral meaning, which is that the weight of 20 means that each of the 250 households selected represented 20 households in the population. This would be important to recognize for the purpose of estimating, say, total expenditures across households — for that we would need to know that each selected household in the sample represented 20 in the population and that the sample total expenditures would need to be multiplied by 20 to get an unbiased estimate of total household expenditures across the population.

Finally, before moving on we focus on one other curious result. The actual variance from equal probability of selection was .39992835, while that from unequal probability of selection but no weights applied to estimation of the population average from the sample was .35071568, leading to the design effect

$$\frac{V^{PPS,NW}}{V^{EPSEM}} = \frac{.35071568}{.39992835} = .87694628$$

In other words, we *actually did better* (from a sampling variation standpoint) by engaging in unequal selection probabilities but then ignoring weights than we might have with equal selection probabilities.

Although this might seem counterintuitive (given our findings about the consequences for sampling variances of using weights), there is a fairly simple explanation — there really can be less variation in the mix individuals in from sample to sample with unequal selection probabilities. To begin with, recognize that we are simulating 1,000 selections of samples of size 250. Thus, across the simulations there are 250,000 selections of households.

We might expect that the average household would be selected 50 times ($= .05 \cdot 1,000$). Under equal probability of selection the number of times a given household was selected ranged between 26 and 78 times, with the roughly 80 percent of households in the 10^{th} to 90^{th} percentiles of the distribution of the number of selections (i.e., the number of times each household was selected) picked between 41 and 59 times. In other words, the distribution of the number of times that the various households in the 5,000 household sample were picked is pretty tightly packed around the expected mean of selection 50 times.

This is intuitively expected. Because households were selected with equal probabilities of selection, no households had a persistent advantage or disadvantage in terms of selection probability. Thus, we would have expected a fairly tight distribution of the total number of times each household was selected around the expected mean value of 50.

This is illustrated in Figure 2.11. Figure 2.11 shows the cumulative proportion of the 250,000 selections during the 1,000 sample selection simulations across households in the population when those households are ranked from the least to the most selected. The equal probability of selection case is represented by the red (upper) curve, which is close to a 45-degree line, indicating that each household's contribution to the cumulative number of selections is not wildly out of line with the case where each household makes roughly the same contribution (that is, each household was selected roughly the same number of times).

Unequal probability of selection is very different. In this case some households had much greater probabilities of selection than others. The distribution of the total number of times each household was selected shows nowhere near the concentration of the equal probability of selection case. In both cases the mean number of times the household was selected was 50. In the unequal selection probability case, the minimum number of times a household was selected in the 1,000 simulated sample selections was 0 and the maximum number of times one was selected was 120. The roughly 80 percent of the distribution between the 10^{th} and 90^{th} percentiles of the distribution being picked between 11 and 89 times. But this also means that the roughly 10 percent of the sample picked most often was picked between 90 and 120 times. In fact, the households in the most selected 10 percent of the population contributed roughly 18.75 percent of the 250,000 household selections across the 1,000 simulated samples! By contrast, the least selected 10 percent contributed about 1.35 percent of the 250,000 selections across the sample selections.

This is illustrated by the blue (lower) curve in Figure 2.11. Now, the cumulative contribution of households grows much more slowly (as the households that contributed the fewest times to the

250,000 selections were selected few times), before becoming steeply upward sloping (representing the large number of times the most selected households were selected).

The point is that, compared with the equal probability of selection case, the same households keep showing up again and again in the samples selected with unequal selection probabilities. The comparatively recurring set of households across samples dampens sample by sample variation in the estimates of the population average daily expenditures. It is also worth noting that the comparatively (to the equal probability of selection case) low sampling variability of the estimates with no weights but unequal probability of selection brings into high relief just how bad the consequences of applying weights really can be; they can worsen sampling variation not just compared with the equal probability of selection case but even compared with a *more* efficient possibility *based on the same selections as the weighted estimates.*

One might be tempted to think that selection with unequal probabilities of selection and then no application of weights is a path to fairly efficient (low sampling variation) estimates. However, we must recall that this is a case of uninformative weights, and so we effectively got away with ignoring them in the unequal selection probability case as they did not effect the estimate of average

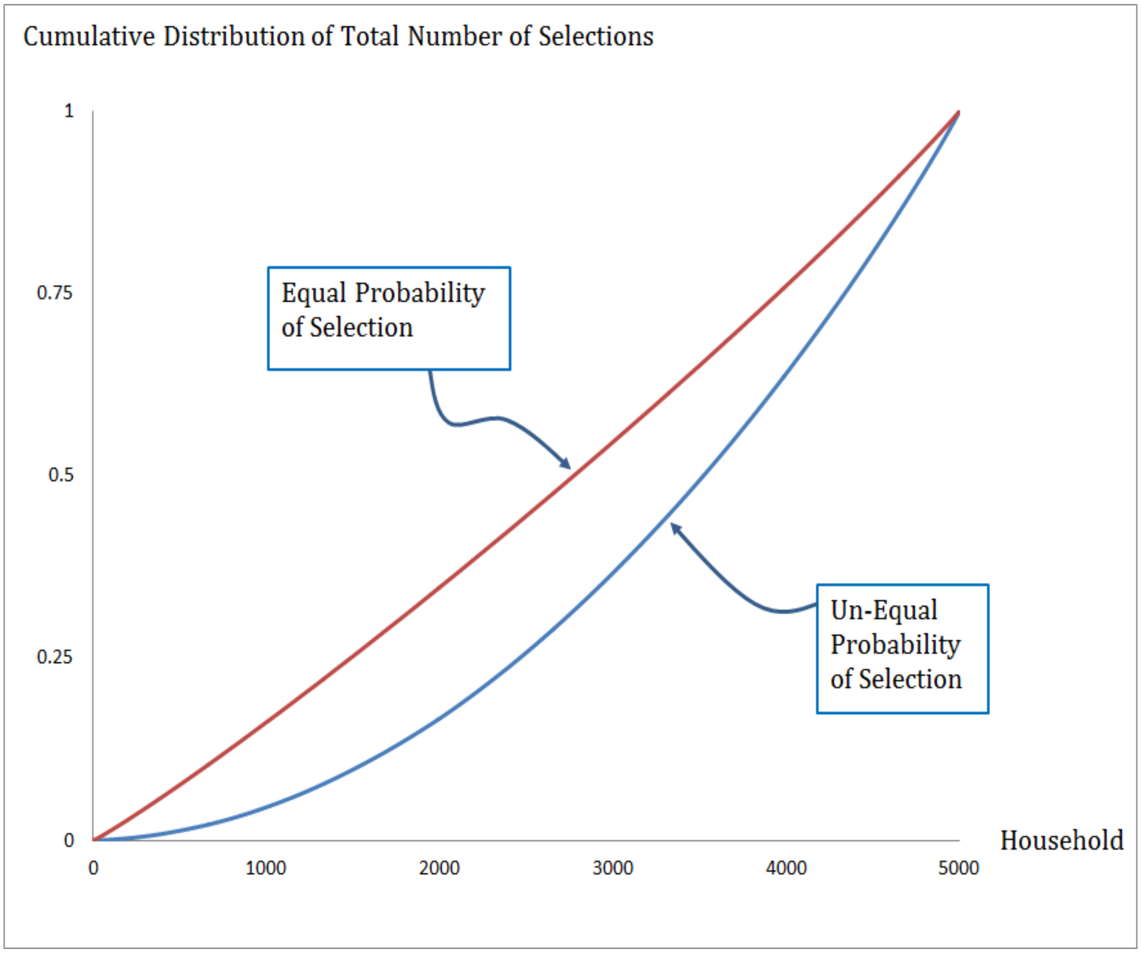


Figure 2.11. The representation of each individual under equal and unequal probability selection

expenditures.

This brings us to **Departure 4**, in which we consider informative weights. Our simulations for this departure are based on STATA code Chap2-6.do. To begin with, we simulate our reference population of 5,000 households in the same fashion as in the earlier examples, including their original expenditure value (hence the true "population" mean is still 39.29789 across the population of 5,000 households). We then simulate a population (number of members) for each household. To do so, create a variable v that is expenditures Y_i plus 9 times a draw from the standard normal distribution, $v = Y_i + 9 * \mu$ where μ is the draw for each household from the standard normal distribution. Finally, we create a new variable *pop* that is the 20 quantiles of v.[32] The result is a variable *pop* that runs from $1, ..., 20$ and represents household "population." Because this variable is based in part on Y_i it should be related to that original household expenditure variable.

The reason why we created the variable $inc2$ in the fashion that we did in Departure 3 can now be made a bit clearer. The variable $inc2$ was created in the same fashion algorithmically as Y_i. This was part of an attempt to make the generation of the population in Departures 3 and 4 as similar as possible, so that any differences in performance could be ascribed to the greatest extent possible the one way in which the two simulation exercises clearly *do* differ: The household population measure is now correlated with household expenditures.

The resulting population is one in which household expenditures and population have a correlation of 0.7293. This is a very high correlation but not a behaviorally implausible one — larger households (in terms of the number or population of residents) probably *should* have higher daily expenditures. By contrast, in the Departure 3 simulation the correlation between household expenditures and population was just 0.0046.

Having generated the population of 5,000 households with their expenditures and population values, we then repeat the following sample selection simulation 1,000 times:

1. Select 250 households by equal probability of selection (epsem) and estimate population daily expenditures across the sample by simple average of daily expenditures for the selected households.

2. Select 250 households by probability proportional to size (pps) sampling, with the household population as the size measure.

3. Compute the simple unweighted average for daily expenditures across this new sample of 250 households.

4. Calculate the *ex-ante* sampling probability for each selected household i as

$$p_i = 250 \cdot (pop_i/POP)$$

[32] In other words, we sort v in ascending order, assign each value for v to equal sized bins, and then create a new variable *pop* such that smallest 250 values for v are in bin number 1, the next smallest 250 values for v are in bin number 2, etc. For each household, *pop* is then equal to the bin number for that households draw for v.

where pop_i is the population of household i and POP is the total population size of the frame (i.e., $\sum_{j=1}^{5000} pop_j$).

5. Calculate the sampling weight for each selected households as

$$w_i = \frac{1}{p_i}$$

. Higher probability households thus have a lower weight.

6. Using this weight calculate the weighted average

$$\frac{\sum_{i=1}^{250} w_i \cdot Y_i}{\sum_{i=1}^{250} w_i}$$

for each of the $i = 1, ..., 250$ households selected for inclusion into the new sample.

In other words, we basically repeat the simulation for Departure 3, this time with the household population correlated with household expenditures.

Against the true population average daily expenditures of 39.29789, the average of the estimates of the same across the 1,000 simulated sample selection exercises were 39.301138 (for the estimation of population average expenditures by simple sample averages after epsem selection), 43.375641 (from the estimation of the population average by simple unweighted averages after selection with unequal probability of selection) and 39.349694 (from weighted estimation after selection with unequal probability of selection). Now we see the problem with simply ignoring weights in the unequal selection probability case: when those weights are informative, the result of ignoring them is a biased estimate. On the other hand, weighted estimation yields an unbiased estimate.

To make sure it is absolutely clear before moving on, we expand a bit on why these are informative weights in the pps selection case. We have already noted that the correlation between daily expenditures and household population of 0.7293. Selection with pps involves assigning a higher probability of selection to households with more members. However, since larger households tend to have higher expenditures, this means that households with higher expenditures have a greater probability of selection (compared with their actual representation in the population).

In fact, across the population the correlation between the probability of selection

$$p_i = 250 \cdot (pop_i/POP)$$

and household expenditures is also 0.7293. Essentially, this means that our selection method is slanted toward selecting higher expenditure households. Consequently, across the 1,000 simulated samples, 38.3056 percent of the households selected came from the highest expenditure 25 percent of households, while those in the top half of households by expenditure represented 66.6496 percent of selections. By contrast, households in the bottom 25 percent by expenditures constituted just 12.0632 percent of the households selected across the 1,000 simulations. In short, higher expenditure households are over-represented in our selections compared with their actual population shares.

The weights applied are simply the inverse of probability of selection:

$$w_i = \frac{1}{p_i}$$

Higher probability households thus carry a lower weight in estimation. But this means that the weights are also correlated with expenditures (in our simulation to the tune of -0.5786).

A fairly clear story thus emerges. Because expenditures and household population are positively correlated, pps sampling based on household population means that selection is slanted toward higher expenditure households (where the “slant” is relative to their actual share in the population). The resulting sample has a larger proportion of high expenditure households than their actual presence in the population. Estimation of expenditures by a simple (i.e., unweighted) average thus leads to an over-estimate of population average household expenditures. The weights are negatively correlated with expenditures. The application of weights dampens the information contribution of higher expenditure households to the estimate of the average (and amplifies the contribution of lower expenditure households, which are under-represented in the sample), bending the information contribution of households back toward their actual share in the population.

The fact that weights provide a clear benefit in the informative case in terms of delivering unbiased estimates does not, however, mean that we escape the efficiency consequences of them. The average of the estimates of the design effect in the unequal probability of selection estimation with weights case via `estat effects` is 2.6527494, while the average for the two Kish-inspired estimators of the design effect were around 1.85. The design effect

$$\frac{V^{PPS,W}}{V^{EPSEM}}$$

calculated from actual sampling variation across the 1,000 sample simulations is 2.8176274.

The comparatively low estimates of the design effect from the Kish estimators discussed earlier are at first glance a curious result. The Kish estimators are based on what are usually referred to as “arbitrary” weights, which means uninformative (i.e., not correlated with the outcome). There is no one “formula” (at least that we know of) for the informative weights case. We have seen approximations more or less along the lines of

$$var\left(\frac{\sum_{i=1}^{n} w_i \cdot Y_i}{\sum_{i=1}^{n} w_i}\right)$$

$$\approx \frac{1}{\left(\sum_{i=1}^{n} w_i\right)^2} \cdot \left[var\left(\sum_{i=1}^{n} w_i \cdot Y_i\right) + r^2 \cdot var\left(\sum_{i=1}^{n} w_i\right) - 2 \cdot r \cdot cov\left(\sum_{i=1}^{n} w_i \cdot Y_i, \sum_{i=1}^{n} w_i\right)\right]$$

where r is

$$\frac{\sum_{i=1}^{n} w_i \cdot Y_i}{\sum_{i=1}^{n} w_i}$$

This is an approximation and, as such, it unsurprisingly performs better in some circumstances than others. This formula suggests that a positive correlation between the outcome and the weights can actually mitigate the variance consequences of weights. On the other hand, given the negative correlation between the weights and the outcome in our example, the elevated sampling variance is less surprising.[33]

The big "take-away" from Departure 4 is that use of appropriate weights (i.e, weights that are the inverse of the probability of selection) can make unbiased estimation of population parameters possible when selection is based on unequal and non-randomly determined probabilities of selection. That said, it is worth asking if there are conditions that need to be met to construct weights that will achieve this.

While we could discuss this for a very long time with many unusual cases, twists, and turns, there is an overarching broad truth: Appropriate weights can be crafted when selection conforms to the requirements for probability sampling. In short, probability sampling requires that each member of the population have a positive, known probability of selection. The first precludes the possibility that any members of the population are excluded from selection (we have seen the problems that can cause in the discussion of Departures 1 and 2). The second insures that we can always craft the probability weight

$$w_i = \frac{1}{p_i}$$

for each of the $i = 1, ..., N$ members of the population, should they be selected.

Probability sampling is not guaranteed, and indeed there are many alternative selection mechanisms typically lumped together under the label non-probability sampling. These methods have in common that they violate in one fashion or another the requirements for probability sampling.

Consider several examples. First, convenience sampling involves selection from a portion of the population easy to select or close at hand. The classic example is the opinion surveys conducted by selecting passers-by at shopping malls or other public venues. There are two immediate problems with this approach to selection from a probability sampling standpoint. First, some members may have no probability of selection (they may never visit the mall during the period when the survey is being conducted). Second, it is generally impossible to know the probability of selection of those selected, and thus whether their selection probability exceeds or falls short of the average for the population of interest.

Another example is so-called "snowball" sampling. Although there are many variations on this, the common thread is that one selection somehow leads to the next (perhaps through referral). It should be clear that this does not allow the assessment of the probability of selection. Less clear is whether it excludes some members of the population.[34]

[33] There instances where weights can actually lower variance compared with not using weights. For instance, this can happen under the so-called Neyman allocation, which we do not discuss since there is rarely the opportunity to use it in international work given the lack of appropriate, accurate prior information.

[34] In the United States, there used to be a popular party game called the "Kevin Bacon" game. It was based on the logic that no one in the United States is more than six degrees of separation (six links of personal acquaintance) from the popular U.S. actor Kevin Bacon. Other variants attempted to establish the participant's degrees of separation from other public figures. These games were all motivated by the (completely unproven) suggestion that all people on earth are at most six degrees of separation from each other. If this is true, then in fact in theory many variants of snowball sampling where six or more respondents are selected could in theory lead to the selection of anyone on earth. Of course, one still would not be able to determine their true probability of selection, since it would be based

To make clear that the distinction between probability and non-probability sampling is more than just an abstract consideration, we discuss one of the most epic failures in the history of survey research: the misprediction of the 1948 U.S. election by political polls. Political polls are simply surveys to determine political leanings, in this case of presidential election polls likely voting patterns in upcoming elections. In 1948, pollsters predicted that Thomas Dewey would defeat Harry Truman, prompting one newspaper to actually print the wrong headline. This lead to the iconic photograph of President Truman in Figure 2.12.

Essentially, the polls failed in 1948 because the samples collected for them were selected by non-probability sampling. One manifestation of this common to nearly every poll at the time was the use of a species of non-probability selection known as "quota sampling." Under quota sampling, interviewers are assigned quotas of respondents exhibiting certain characteristics (by sex, race, income, and other characteristics) to interview regarding who they supported in the presidential election. Clearly, assignment can only occur in terms of characteristics the interviewer can readily identify in respondents.

Introduced to political polling in 1935, quota sampling had been used since the 1936 presidential election. Some polling firms, such as Gallup, relied on face to face interviews, with interviewers sent out with quotas for different types of individuals to interview. There were several major problems with this in practice. First, in their attempt to meet their quota for each type, interviewers may effectively have resorted to convenience sampling or other types of less than fully random selection. Second, the quotas assumed that the distribution of types in terms of the characteristics considered was known for the population of interest (and the population of interest was really those who would vote on election day). If, of course, the polling companies had guessed incorrectly about the composition of the population of voters on election day, then the quotas would produce a population profile that did not match that of the population of actual voters. Finally, and perhaps most importantly, quota sampling involved a pre-determined set of characteristics, and there is always the possibility that a key characteristic correlated with voting behavior had not been considered in setting the quotas.

Distinct from the selection methods and logic, the methods of identification and enumeration of potential survey respondents may have introduced bias. In other words, the ways that the sampling frames (lists from which individuals were selected for interview) were constructed may have provided a skewed launching point for selection (compared with the actual distribution of types of individuals within the population). For instance, it was not uncommon for polling companies to draw up lists of potential respondents from phone books, automobile registration lists, etc. The trouble with this is that behavioral choices (such as owning a phone or a car) were determining eligibility for selection into samples in a non-random fashion (sound familiar?).

The result of these practices was a general failure by polling firms such as Gallup to predict correctly Dewey's vote shares. Problems first arose clearly in the 1936 election but did not change the overall prediction of the eventual winner until the 1948 fiasco, when polling firms generally predicted a comfortable victory by Thomas Dewey over incumbent President Harry Truman. Truman actually

on poorly understood, non-random mechanisms such as referral.

defeated Dewey, shocking pundits and leaving at least one newspaper (the Chicago Daily Tribune) an embarrassing role in an iconic photo.

The lesson of the 1948 fiasco is that correctly conducting probability sampling really matters. Non-probability approaches, no matter how earnest one is in pursuing them, as essentially never guaranteed to deliver reliable inferences about populations.

It is easy to assume that modern survey practice precludes such failures, consigning them to the increasingly distant past. Unfortunately, this is not the case. Recent elections (such as the Greek vote on E.U. austerity measures in 2015 and the British parliamentary elections of 2014) have seen widespread failures of polls to predict outcomes even reasonably correctly. The need for vigilance and careful thinking about first principles of sampling is as urgent now as in 1948.

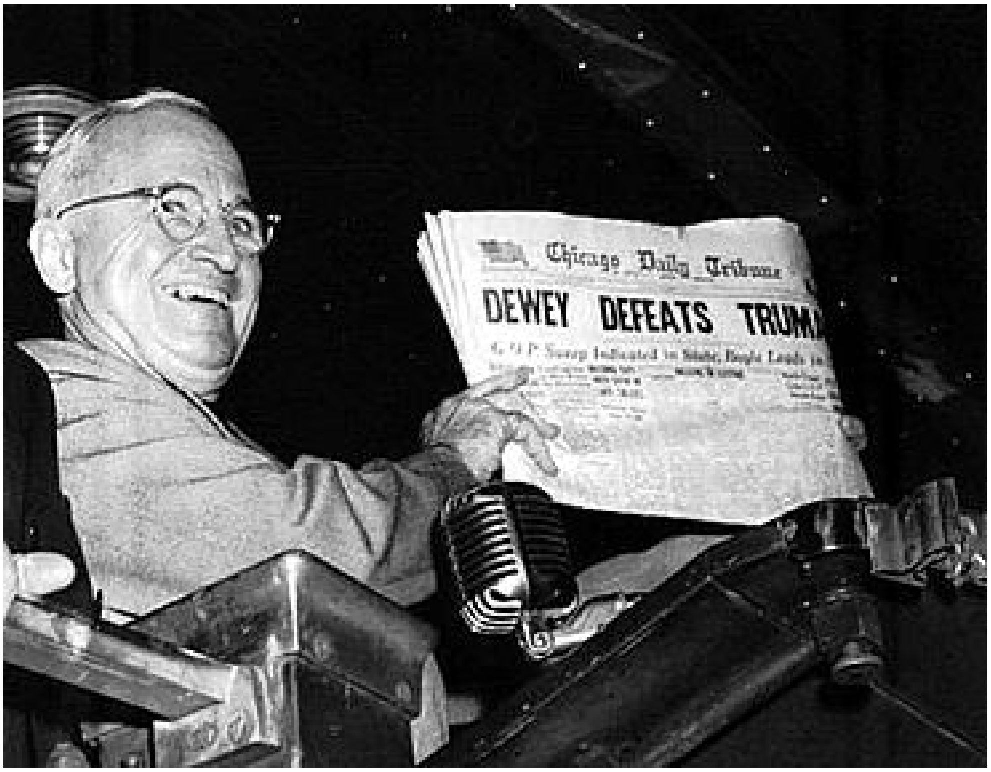


Figure 2.12. Non-probability sampling has consequences

To summarize the discussion to this point, we have learned that:

- Population parameters can be learned from samples gathered from those populations
- Selection of the samples across all members of the population with equal probabilities of selection yields self-weighting samples that can support unbiased and comparatively efficient estimation of population parameters
- Failure to select across all members of the population of interest can lead to biased estimates of parameters for that population except in the extraordinary case where members of that population are excluded from selection at random

- Unequal probabilities of selection can be addressed with sampling weights based on the probability of selection
- These weights will make a difference on average to estimates only if the weights are informative (i.e., correlated with the outcome of interest)
- There is an efficiency cost to unequal weights
- In the event of unequal selection probabilities, weights can be crafted to generate unbiased estimates as long as selection conformed to the requirements of probability sampling, which requires that all members of the population have a known, positive probability of selection

These considerations are universal to population surveys, including those geared toward impact evaluations. Surprisingly often we have seen designs for impact evaluation surveys floated that failed to take into account on or more elements of this list. We will now turn to some complications in the sample selection process that must be taken into account when thinking about sampling for impact evaluations or any population survey objective.

2.2 Common Sample Selection Extensions and Complications

Having introduced simple, independent sampling from a single frame that lists an entire population of interest, we now turn to some extensions and complications that arise frequently (indeed, our first instinct was to write "constantly") during sample selection in population survey research. Some have implications for design effects.

2.2.1 Multistage Selection

Up to this point, our motivating example has been one in which we select a sample of our unit of analysis (households, in our running numerical example) directly from a frame containing the complete list of those units of analysis. The use of a household as our hypothetical unit of analysis was altogether fitting for a manual focused on impact evaluation, as units of analysis in impact evaluation tend to be rather low level or "micro" (such as individuals or households).

However, typically we will not have a complete list or frame for sampling units as low-level or granular as individuals or households. Even in wealthy, post-industrial societies with their rich sample frame infrastructures (translation — they have many lists of all sorts of things that could serve as frames for all sorts of surveys) a list of individuals or households would generally not be available at a national or even large subnational level. This presents an immediate complication for selection.

The usual solution is what is known as "multistage" selection. Under the simplest version of this approach, we first select a larger sampling unit that contains many units of analysis. Having selected that larger unit, we then list all of the units of analysis in that unit and select from that list. Because there are two steps to selection in this simple case (select from the list of larger units then, within each selected larger unit, select units of analysis), this is a case of two-stage selection.

For instance, if we wish to have a sample of households, we might first select census clusters (perhaps from a publicly available national frame of census clusters) and then list and select households within selected clusters. The first unit chosen (in this case the census cluster) is called the "primary" sampling unit. The next unit selected (households, in this example) is the "secondary" sampling unit.

Sometimes multistage selection can involve quite a few stages of selection of larger (where by "larger" we really mean "containing multiple units of analysis") sampling units. For instance, we might first select a higher level sampling unit. Having done that, we might then select a somewhat lower level sampling unit. Typically, this means that in each selected higher level sampling unit, we might select one or more lower level sampling units. This process continues until we get down to the level of the unit of analysis. For instance, to gather a sample representative of the individuals in the urban areas of some nation, we might first select cities, then wards within selected cities, then blocks within selected wards, then households within selected blocks and finally individuals within selected households. In this example, cities are the primary sampling unit, wards are the secondary sampling unit, blocks are the tertiary sampling units and households are the penultimate sampling units.

The major advantage of multistage sampling from the selection standpoint is that it essentially reduces the dimensionality of the challenge of developing a list of units of analysis from which to select. For instance, building a list of all the households in a society is a prohibitively big challenge. Developing lists of households only in a few hundred clusters is a much more reasonable challenge.[35]

Multistage sampling can satisfy probability sampling as long as there is some path of selection through which every unit in the population represented by the units of analysis has some *ex-ante* (i.e., before the primary sampling unit selections are made) chance of selection. Because the higher level sampling units contain many units of observation, those units tend to be associated with a geographical areas and, when this is the case, are commonly referred to as "area frames."

For instance, a list of wards is a list of administrative areas, but those administrative areas are defined by the physical space that they cover; a ward is an administrative area covering a certain physical/geographical part of a city. When thinking about higher level area frames that will support multistage selection, a commonly invoked phrase is that the sampling units in that higher level area frame should be "mutually exclusive and collectively exhaustive."

Consider the selection of a sample of households from a city. We want to learn something about the population of households in that city. In the first stage of selection of households we will select a sample of wards from a list of the wards in that city.

Since our interest is the population of households in the city, it is important per probability sampling that every household in the city have some positive probability of selection. This is where the "exhaustive" part comes in: When we say that the ward frame is exhaustive we mean that every portion of the physical space in the city is captured by some ward in the frame. Because this is true, every household in the city is in some ward in the frame of wards. Therefore, when we select

[35] From the *fieldwork* standpoint, the major advantage of multistage sampling is that it geographically concentrates samples.

wards, every household in the city is in play in the sense that the ward in which they reside could be selected.

If wards overlapped (i.e., if they were not “mutually exclusive”), then some households would have multiple pathways to selection: they could in principle be selected if *any* of the wards in which their household is found were selected. This introduces the potentially annoying complication of overlapping frame selection (which, again, we will discuss below).

It is in the setting of multistage selection that the potential advantages of pps selection are perhaps clearest. Consider a two stage sampling process. In the first stage, we select a sample of c clusters from a frame of C clusters. The frame includes the household population (household size) of each cluster, h_i (where i indicates the i^{th} cluster in the frame). There are a total of H households across the C clusters in the frame. We then list the households in each selected cluster and select h households from the household list/frame in each cluster.

Suppose we select the clusters by pps sampling. The selection probability for cluster i is

$$P_i = \frac{c \cdot h_i}{H}$$

We then list the h_i households in that cluster and select h of them. The selection probability for household j *within* cluster i is

$$P_j = \frac{h}{h_i}$$

The overall probability of selection for household i in cluster j is the product of these two probabilities:

$$P_{ij} = P_i \cdot P_j = \left(\frac{c \cdot h_i}{H}\right) \cdot \left(\frac{h}{h_i}\right) = \frac{c \cdot h}{H}$$

Notice that this probability is the same regardless of i and j. In other words, using pps sampling in the first stage of selection and epsem sampling in the second has left us with an overall selection probability that is the same across all households in society (and across all households that are selected).

In other words, the probability of selection is constant across households, including households selected. However, the weight is still simply the inverse of the probability of selection:

$$w_{ij} = \frac{1}{P_{ij}} = \frac{1}{\frac{c \cdot h}{H}} = \frac{H}{c \cdot h}$$

Thus the weight does not depend on i or j. In other words, the weight is the same across all potential selected households. This means that the sample is self-weighting and we are, despite using pps selection in the first stage, in a situation very much akin to that from the epsem selection case in Departures 3 and 4 in the last section. Thus, sampling weights would not be required for things like estimation of averages (for the purpose of, for instance, worrying about the unbiasedness of estimates), allowing us to avoid the variance consequences of weights.

This happy state of affairs would not have arisen had clusters been selected via epsem sampling.

For instance, the first stage probability of selection would then have been

$$P_i = \frac{c}{C}$$

leading to an overall probability of selection of

$$P_{ij} = P_i \cdot P_j = \left(\frac{c}{C}\right) \cdot \left(\frac{h}{h_i}\right) = \frac{c \cdot h}{C \cdot h_i}$$

and a weight of

$$w_{ij} = \frac{C \cdot h_i}{c \cdot h}$$

This weight depends on i. Hence we would have sampling weights that differed between households selected from different clusters. To the extent that this was informative, we need to use these weights in estimation of statistics like averages, with consequences for sampling variance.

If a size measure had not been available for clusters, one possible way to achieve a self-weighting sample might have been to select a certain constant percentage of households in each cluster (as opposed to the absolute target h). This would have led to a weight of

$$w_{ij} = \frac{C}{c \cdot \phi}$$

where ϕ is the proportion of households to be selected in each cluster.

While superficially appealing, this solution has some practical drawbacks. First, if you do not know the total number of households across clusters, it is very hard to set an *ex-ante* (i.e., before selection) value of ϕ that achieves the household sample size target. This often leads to the need to either update ϕ as listing in clusters proceeds (which leads us right back to unequal selection probabilities and weights) or settle for under- or over-shooting the sample size target. Second, there are some theoretical complications to this. For example, if the final household sample size is not clear from the outset (in other words, if the sample size is a variable), then establishing unbiasedness becomes a lot more complicated theoretically (though in many instances this is much ado about nothing in practice).

A practical consequence of multistage sampling can be elevated sampling variances via a different avenue called clustering. To fix ideas, assume that the multistage selection yields a self-weighting sample (allowing us to abstract away from the sampling variation consequences of variable weights). Further, we continue our example of selecting clusters and then households within clusters.

If we had a list of households across the population, we could select households directly. The result would be selections that were independent (in that the selection of one household was not related to any event that had any influence on the selection of another household) scattered throughout the population. It would be reasonable to view the outcomes (for instance expenditures) across all households as completely independent outcomes.

This is not true in multistage selection. In that case, the selected households would not be scattered evenly through the population but instead concentrated in the selected higher level sampling units. The unit above the household is commonly referred to as the cluster (though the term has sometimes been applied in other ways). In multistage selection there would be households whose selection was determined by the same event, namely the selection of the cluster they share.

This also means that they share any cluster-level factors (culture, access to markets, social institutions, etc.) that might influence outcomes such as expenditures. Two households drawn from the same cluster would thus yield less true information toward the estimate than two households that were selected completely independently. To a certain extent the two households from the same cluster are in a sense providing repetitive information because part of what they reveal about the outcome of interest (e.g., expenditures) is driven by the same factors. Put slightly differently, by concentrating households in a (typically small) proportion of clusters, we limit the variation available given what would have emerged from a truly independently selected sample that exhibited a fuller or richer range of variation in the circumstances in cluster level factors that influence the outcome.

Despite this efficiency cost, multistage selection remains a popular method in population survey work, particularly in the development context. The reason is simple: The cost of elevating samples to account for the efficiency price of clustering is still typically far less than the cost of crafting a list of units of analysis across the population of interest.

2.2.2 Stratification

Another twist on sample selection is stratification. Basically, stratification involves dividing the population into subpopulations and then selecting samples separately from those subpopulations. The usual reason for this is to insure sufficiently large samples from population subgroups. For instance, in population surveys in virtually any society with any meaningful ethnic or racial heterogeneity, it is common practice to select separately samples from smaller minorities. The idea behind this is that, given their small share of society, were these minorities to be selected as part of probability samples selected across the entire population we would likely end up with empirical samples from these populations too small to form estimates of any meaningful precision for them. For this reason, it is often said that a population survey's sampling design "oversamples" such subpopulations.

Viewed from a slightly different standpoint, stratification is an economical strategy. Suppose that we have some parameter for which we suspect we will need about 10,000 observations to estimate with sufficient precision in any given population (with more on that issue in the next chapter). Suppose as well that we are interested in obtaining such estimates for group A (who constitute 90 percent of the population) and group B (who constitute 10 percent of the population). If we select 10,000 sampling units from across the population, we would expect around 9,000 to come from group A and 1,000 to come from group B. In other words, we would not have enough sample size for either subpopulation to estimate the parameter with the precision we want. If we select a sample from across the population of a size large enough (11,111 observations) to obtain a sample of 10,000 observations from group A, we still have far too few (1,111) from group B to meet our objectives in terms of the precision of the parameter estimate. If we select a sample from across the population of a size large enough to insure a sample of sufficient size from group B, we'll end up with a sample size of 90,000 for group A, about 80,000 more than we need (a colossal waste of resources).

On the other hand, if we stratify, we can select samples separately from the populations for groups A and B. This will lead sample sizes of 10,000 for each or 20,000 in total. This allows us to meet the objective for both groups (to estimate the parameter with the desired degree of precision) while avoiding the colossal waste associated with selecting a sample across the population that will be expected to yield a subsample from group B large enough to meet that objective.

On one level, stratification is nearly ubiquitous in sampling for surveys to estimate program impact. We almost always want samples of participants and non-participants (and, as we will see in the next chapter, of minimum sizes in both cases). Beyond that, it is not uncommon for impact among certain population subgroups to be of particular interest. Stratification may be required to insure that the sample sizes obtained for these subgroups are large enough to estimate impact among them.

In and of itself, stratification tends to lower sampling variation (and hence, other things being equal, leads to smaller design effects) for estimates *across* subpopulations that were the focus of stratified sampling. To be clear, by this we mean estimates for the entire population or combinations of subpopulations from which we selected by stratified sample. For instance, if we stratify by gender, the lower sampling variance would be for estimates across males and females.

The reason for this is that stratification tends to eliminate one driver of sampling variation in those overall population estimates — variation in the shares of various subpopulations across samples. Consider, for instance, a survey for which we intend to produce population-level estimates of parameters as well as sex-disaggregated estimates of those parameters. From sample to sample, there will be some variation in the shares of the sample for men and women. By stratifying by gender, we shut down this pathway for variation in the estimate of the overall population-level parameter.

Of course, this will only reduce sampling variation in parameter estimates for the overall population or combinations of stratified subpopulations if the value of the parameter varies across the subpopulations over which we stratify. Consider, for instance, attitudes toward domestic violence by men and women. Stratifying sample selection by gender will only reduce sampling variation in our estimates of overall population level attitudes toward domestic violence if these attitudes varied on average by gender. In that case, sampling variation in the shares of each gender across samples should contribute to sampling variation in the estimates of population attitudes of toward domestic violence. Of course, if these attitudes did not vary by gender, stratification will not help reduce sampling variation in the population-level estimates of attitudes toward domestic violence.

Of course, stratification could in principle introduce a variance inflating force: weight variation. Because stratification tends to lead to oversamples for some groups relative to their shares in the population, the members of these groups have a higher probability of selection and hence will earn a lower weight in estimation (of parameters across the population) than the rest of the population.

In other words, stratification could, in principle, introduce or worsen weight variation.

One obstacle to stratification in many applications is the lack of information to support stratification. If you wish to stratify selection by gender, ethnicity, race, socioeconomic status, or any other potentially meaningful characteristic, you have to be able to parse your overall frame according to these characteristics. Many times the frame won't contain this information. Where interest particularly lies with one subgroup not identifiable from readily available frames, one might need to resort to one or more of the possibilities discussed in the subchapter below dedicated to "Hard to Find Subpopulations."

2.2.3 The Design Effect, Re-visited

We have now met the three big sample selection factors that shape the design effect. As a gross generalization, we can state the following:

- The design effect is increased by increased weight variation.
- The design effect is increased by clustering. This has two facets:
  - The design effect will tend to be increased by concentrating the ultimate sampling units among fewer and fewer clusters.
  - The design effect will tend to be increased as the intra-cluster correlation (crudely, the intra-cluster correlation is a measure of the similarity of sampling units within clusters; it is typically expressed as a comparison of the within-cluster variance with the between-group variance) of the outcome of interest increases.
- The design effect tends to be reduced by stratification, though it can be worsened to the extent that stratification is driving weight variation.

As a gross generalization (and we do mean gross, as there are many bright folks who would object to what follows), we have found that the first two effects are stronger in typical practice. One way of thinking about this is that one should probably not bank on stratification offsetting the variance consequences of clustering and unequal selection probabilities.

2.2.4 Hard to Find Subpopulations

Sometimes we are interested in a subpopulation and conventional stratification is not an option, typically because the subpopulation is not identifiable on the frames immediately or readily available for sample selection. This is a difficult situation for which there is sometimes no good fix. However, there are a few possibilities one can consider.

One is what the authors sometimes casually refer to as "sort sampling." The basic idea of this is to select a sample from the overall population (via, presumably, the conventional or most convenient frame, which happens to lack an identifier for the subpopulation[s] of interest). The selected units are then subjected to a rapid assessment instrument to decide whether they are members of the subpopulation of interest. Those that are not are either discarded or, in the event where they are of some interest as well, selected at perhaps a different rate than those in the subpopulation of interest (this then becomes a kind of quasi-stratification).

Typically this requires oversampling. Suppose we are interested in a sample of women from an ethnic minority in some society. Assume that in this society households tend to be ethnically homogenous (i.e., households tend to be entirely composed of one ethnic group), households from the ethnic minority of interest constitute probably 10 percent of all households, and there tends to be an average of 1 woman per household. Finally, assume that we cannot identify ethnic minority status from the national census frame of households (in most societies it is unlikely that a census frame of households exists, but let's run with the idea for the sake of simplicity in our present example).

If we needed a sample of 10,000 women from this ethnic minority, one possibility would be to select 100,000 households from the national frame of households. We then do a rapid assessment of all selected households to determine the ethnicity of their members.

There is a trade-off inherent to this strategy. Where subpopulation membership is not immediately obvious (for instance, what if interest lies with poor or undernourished households?) one can face a stark choice. On the one hand, the initial rapid assessment instrument can be kept brief. This carries the consequence of potential mis-assignment of households to the sample we wish to select from subpopulation. For instance, households not truly from the subpopulation may end up in the eventual sample for that subpopulation, while others who did belong are excluded. To the extent that this occurs, the selected sample is potentially less and less representative of the subpopulation of interest (and indeed it can become hard to know precisely the subpopulation for which that sample is representative).

On the other hand, the most obvious remedy for this (a more precise and exacting, and hence presumably accurate, rapid assessment instrument) is fairly costly. For instance, if the rapid assessment tool gets long enough, then in the above example, at some point you are essentially conducting a survey of 100,000 households from a cost standpoint.

To the extent that there is geographic variation in the presence of the subpopulation of interest, one intriguing possibility is to build a geographic information system (GIS) to achieve more efficient sampling from otherwise hard to find or identify subpopulations. This is the subject of a recent manual on GIS and sampling (Lance, Spencer, & Hattori, 2014). The basic idea is that GIS can be used to identify areas where the subpopulation of interest is concentrated. Sampling units from these areas can then be over-sampled (for instance, in the context of some multistage sampling process). This should yield larger final sample sizes from the subpopulation of interest.

To fix ideas, we repeat an example from Lance, Spencer, & Hattori (2014). Suppose that we are interested in learning (via a population survey) something about behavior in some city in the setting of a dengue fever outbreak. Dengue fever is an acute febrile illness spread by humans and the *Aedes Aegypti* mosquito, with the cycle for transmission being (more or less) an infected mosquito biting a human. The human is then infected and could infect any mosquitoes that bite them while they are in the throes of dengue infection.

Suppose that sample size estimation has suggested that we need a sample of 112 clusters with a dengue outbreak. To achieve this we will have a budget of $300,000. The plan is to select a sample of clusters, monitor them for an outbreak for the study interval (at a cost of $200 per cluster) and, in the event that an outbreak is detected, conduct survey fieldwork in those clusters (at a cost of $2,000 per cluster). Now, history suggests that around 4.9 percent of clusters will experience a dengue outbreak. This implies an expected cost per cluster of $298 ($\approx 200 + .049 \cdot 2000$). With a $300,000 overall budget this means we would be able to afford to select just 1,006 clusters, of which we might expect 49 ($\approx 1006 \cdot .049$) to have an outbreak, far short of our target of 112.

However, it turns out that the mosquito that transmits dengue (*Aedes Aegypti*) has some interesting habits. It rarely flies more than a couple hundred meters from where it was born. This helps to explain why dengue cases tend to occur in clusters. It is also (comparatively) very picky about where it lays its eggs: it tends to prefer relatively clean water. This helps to explain why dengue is largely a rainy season phenomenon in place like South and Southeast Asia — the rain creates new pools of reasonably clean water. These pools tend to form in certain places (low lying areas, areas with great places for puddling like tire dumps, gardens, and construction sites, etc.). Moreover, the places where dengue has broken out in the past provide a good indication of where it is likely to break out in the future.

All of these are places that can be potentially modelled through a GIS. What we are saying is that there is a geography to dengue. Suppose that we build a GIS that through its layers allows us to sort the city into two types of clusters: low outbreak risk clusters (these are 95 percent of clusters and carry only a 2 percent risk of outbreak during the survey interval) and high risk clusters (5 percent of clusters, these have a 60 percent risk of an outbreak during the study interval).

This allows us to stratify selection by risk. For instance, we could select 180 high-risk clusters (with 108 expected to experience an outbreak) and 200 low-risk clusters (with 4 expected to experience an outbreak). In other words, we might expect a sample of 112 clusters with an outbreak for the same $300,000.

Note that this would still satisfy probability sampling as long every area of the city (and hence every household and individual in that city) resided in either a high- or low-risk area (per the GIS model). Even though most of the sample with an outbreak might come from the high-risk cluster, all clusters (and hence all households and individuals within them) would have a positive probability of selection.

Of course, building a GIS isn't free. However, a basic calculation reveals that achieving a sample of 112 outbreak clusters without such stratification would cost $681,142.86. But one way of looking at this is that we could spend up to $381,142.85 on building the GIS and still be better off doing so in terms of, for example, expenditures per outbreak cluster identified.

This isn't a solution to every problem. For instance, Lance, Spencer, & Hattori (2014) discuss West Nile virus, which is spread by a different mosquito (*Culex*) that generally has a much greater range. Moreover, birds play a key role in spreading West Nile (in fact they are the source of it). Because of this there is much less of a concentrated geography to West Nile risk (the birds and mosquitoes might fly all over the place) and stratification based on GIS risk assessment would be nowhere near as productive.

There are also some subpopulations for which particular sampling strategies have been developed. The sampling statistician William Kalsbeek has, for instance, applied considerable thought and creativity (see, for example, Kalsbeek, 1986, Kalsbeek & Cross, 1982) to the task of sampling nomadic populations. His approach to the problem of sampling from nomadic populations provides a good example of how best to frame thinking about sampling subpopulations. Kalsbeek (1986) begins by defining the subpopulation precisely:

> Pure nomads are those who depend entirely on their animals for livelihood and who, as a result of this dependence, move periodically about in search of water and pasture, thereby having no permanent residence. Semi-nomads are similar to pure nomads except that they also engage in agriculture during part of the year when crops can be grown. However, during the dry season both types of nomads herd their animals in search of water and grazing lands and are therefore referred to collectively in our discussion as nomads. (p. 164)

He then describes the complications of sampling nomadic households, including the tendency of their households to split and reform, their mobile nature, etc. He then proposes and works through the statistical implications of various sampling schemes for nomadic populations.

We mention Kalsbeek because we view his and his collaborators' work on the subject of selecting nomadic populations as an example to be followed for thinking about how to sample from difficult populations. Even if the reader's interest lies in a different subpopulation, this body of work on nomadic populations is still worth reviewing to get some idea of how to organize one's thinking to developing selection strategies.

Another example worth mentioning is the Priorities for Local AIDS Control Efforts (PLACE) methodology for determining where to reach those most at risk of acquiring and transmitting HIV (see Weir, et al., 2005). PLACE is actually in some sense an umbrella term for a number of activities (including survey activities). For the present purposes, it is notable for providing another strong example of how the identification of a non-obvious (from, for instance, the standpoint of a conventional census sampling frame) population might be found. In some sense PLACE can be viewed as having elements of the traditions reflected by both the GIS strategy and the vein of thinking or approach of Kalsbeek's work on nomads.

2.2.5 Large Clusters and Size Sampling

In the first section of this chapter, we learned about systematic sampling and, particularly, about probability proportional to size systematic sampling whereby the probability of selection for a sampling unit depended on some "size" measure associated with it. To recall, if POP_i is the size of the i^{th} unit and POP is the sum of the sizes for the F units in the frame, or

$$\sum_{f=1}^{F} POP_f,$$

then if we select k units the selection probability for that i^{th} sampling unit is

$$\frac{k \cdot POP_i}{POP}$$

Moreover, selected units are separated by a sampling interval I that equals

$$I = \frac{POP}{k}$$

Selection is then a matter of sorting the order of the sampling units, taking the cumulative size down the frame, choosing a random start between 1 and I, selecting the unit whose cumulative size contains that random start, and then selecting the units whose cumulative size contains every I^{th} increment after that.

One possibility we did not discuss is that of a "big" sampling unit. In this context big is easy to define. The i^{th} unit is big if

$$POP_i \geq I$$

If this is the case, the marginal contribution to the cumulative size of the frame made by the i^{th} unit is greater than I. This means, however, that the i^{th} unit must be selected more than once in probability proportional to size systematic sampling.

Let's think about the implications of this for the selection probability of that i^{th} unit. Since the sampling interval is

$$I = \frac{POP}{k}$$

it is also true that

$$k = \frac{POP}{I}$$

However, the selection probability for the i^{th} unit is also

$$\frac{k \cdot POP_i}{POP}$$

Inserting the result for k, we have the selection probability of

$$\frac{POP}{I} \cdot \frac{POP_i}{POP} = \frac{POP_i}{I}$$

However,

$$POP_i \geq I$$

means that the selection probability is greater than 1, which is nonsensical (probabilities must be confined to $[0, 1]$ to be meaningful).

So how does one respond to such big sampling units? One solution is to select them from the frame as per normal practice and simply allow them to be selected more than once (for an example of guidance on best practice for practitioners that suggests just this, see Bierrenbach, 2008). In a multistage selection setting where the units being selected were higher level units (as is usually the case with size sampling) this would typically mean selecting multiple units for unit i at the next stage of selection. Consider selection of wards and then blocks within wards, where the typical plan is to select one block for each ward selected with probability proportional to size at the preceding stage of selection. A "big" ward might be selected three times during the systematic size selection for wards. We would then select three blocks for this big ward rather than just one. For selection probability and weight calculation, each of the three blocks would be assigned a probability of 1 at the ward selection stage.

Another possibility is to make big sampling units their own strata. This is akin to the stratification strategy discussed earlier, only this time it is driven simply by the need to deal with "big" sampling units. Assuming multistage selection, how many lower level units to select per "big" sampling unit so isolated is up to the practitioner. However, avoiding excessive weight variation should probably be one consideration in determining this.

This stratification solution does have a potential catch, however. As one takes big sampling units out of the frame to deal with them separately, the sampling interval will shrink. This means that some of the remaining sampling units could qualify as "big" even if they did not before stratification of the initial "big" clusters. This could lead to an annoying iterative process until one finally arrives at a remaining frame with no more big units. Some (such as Valliant, Dever, & Kreuter, 2013) have suggested best practices for this. Most of the best practice suggestions essentially amount to simply placing some sampling units above a certain size (such as 80 percent of the size of the smallest of the "big" units on the first path through the frame) into their own strata.

2.3 Complications to Weights

We conclude by addressing a few potential complications to weight construction. Most are fairly trivial (from the standpoint of executing modifications to the weights) though the last one discussed (overlapping frames) can be a quite serious complication.

2.3.1 Non-Response Adjustment

Just because a sampling unit has been selected for inclusion in a survey sample does not mean that that sampling unit will end up in that sample. That unit might choose not to participate in the survey. Indeed, the informed consent process is designed in part to seek affirmative permission from subjects to interview them (that is, their affirmative agreement to join the sample) and make clear that participation in the survey is voluntary.[36]

When selected units refuse to join the sample (i.e., participate in the survey) a wedge in terms of representativeness is driven between the sample selected and the actual sample interviewed. We typically term such refusal non-response and describe those who refuse to participate as non-respondents. To the extent that non-response is a non-random event (for instance, to the extent that

[36] The informed consent process is also designed to clarify the potential benefits and risks of participation in the survey study.

non-response is influenced by systematic characteristics such as age, gender, socioeconomic status, etc., or the individual's value for the outcome variable of interest) the selected and interviewed samples are no longer representative of the same populations. Indeed, all that we can say is that the selected sample is representative of the frame population.

Some sort of adjustment for non-response is often made. To begin with, let us define the weight that is the probability of selection from the frame

$$w_i = \frac{1}{P_i}$$

(where P_i is the probability of selection for inclusion in the sample — as opposed to whether sampling units actually agree to participate and thus end up appearing in the sample — per the selection design) as the *design weight*.

It is so called because it reflects the sample selection design. The non-response adjustment essentially involves the modification to the design weight

$$w_i^{NR} = \frac{1}{P_i^R} \cdot w_i$$

where P_i^R is the probability that unit i would agree to be a survey respondent ($1 - P_i^R$ is thus the probability that unit i would be a non-respondent).

This adjustment factor carries with it a compelling intuitive explanation. Compared with their true share in the frame population, sampling units with a high probability of response (i.e., a high value for P^R) will be over-represented in the sample. The adjustment factor

$$\frac{1}{P^R}$$

will be smaller for these units. This lowers their weight somewhat (and hence reduces the information contribution they make to estimates) to account for their over-representation due to high response probability. By similar logic, the adjustment factor will elevate the weight, and hence information contribution to estimates, of households with lower response probabilities because, by dint of those lower response probabilities, they will be under-represented in the sample compared with their true share of the frame population.

How one implements the non-response adjustment

$$\frac{1}{P_i^R}$$

(which is equivalent to how one models the probability of response P_i^R) is not a matter of settled standards of practice and can depend on the information one has available to them. In some surveys, adjustment is pretty simple. For instance, one adjustment sometimes applied (such as in many Demographic and Health Surveys [DHS]-style surveys) is to insert the non-response adjustment at the household level but computed at the cluster level. For instance, for household i in cluster j the

P^R might be modelled as

$$P_{ij}^R = \frac{h_j^R}{h_j}$$

where h_j is the number of household selected in cluster j and the h_j^R is the number of respondent households (i.e., households that agree to participate in the sample). This is a pretty simple estimator of the probability of response. By not considering any systematic factors that might influence non-response probabilities among households within a cluster, it is in essence rooted in a "missing at random" assumption within clusters.

Some have tried to move beyond this by modeling systematic factors X (where X typically represents various factors, such as characteristics of the sampling units) that might influence response probabilities. The simplest approach is to consider the characteristics in X as discrete or categorical variables and define response probabilities for every possible combination of the discrete or categorical values present in the sample.

For instance, if X included gender and marital status (each recorded as binary indicators), we would compute

$$\frac{n^R}{n^s}$$

(where n^s is the number of sampling units selected and n^R is the number that agree to serve as survey respondents) for (assuming each of the following combinations appears in the sample):

1. Married males
2. Un-married males
3. Married females
4. Un-married females

Often referred to as "weight class analysis," this approach assumes that these observed characteristics (gender and marital status) determine who is a survey respondent and that any remaining variation in response probabilities reflects wholly idiosyncratic randomness. It is essentially a "selection on observables" approach to non-response.

As such, it carries all of the limitations of selection on observables approaches as applied to, say, impact evaluation itself. By its structure, it concedes that missingness can be non-random, but then corrects only for those factors driving the response decision that can be observed. This assumption might be a significant weakness from the standpoint of actually getting estimates representative of those selected; in reality, many intrinsically unobservable factors like patience, openness, etc. likely shape the participation decision. To the extent that these unobserved factors are correlated with parameters we wish to estimate, these models will not provide corrections for non-response that will lead to "unbiased" estimates of program parameters.

At the same time, they are often impractical. To operationalize them we must know the characteristics (those that we feel are important to systematically shaping non-response and are correlated with the outcome to be studied) of respondents *and* non-respondents. However, the only thing we typically know about the latter in many cross-sectional surveys (aside from where they live) is that they said “no.”

This is one reason this basic kind of approach is often seen in longitudinal surveys. In longitudinal surveys we interviewed a sample at time $t-1$ and now wish to re-interview them at t. The weight class analysis would in this setting be based on

$$\frac{n_t^R}{n_{t-1}^s}$$

where n_{t-1}^s are the individuals in a given class (i.e., combination of Xs) interviewed at time $t-1$ and, of those individuals, n_t^R is the number who respond at time t. In this setting, non-response can be driven by many things: death between $t-1$ and t; some other personal transition between $t-1$ and t that makes the individual ineligible for continued inclusion in the study; failure to track the unit of observation; and classic refusal at time t. This type of weight class analysis is possible because we sort units of observation into classes based on what we do know about them — that is, their lagged characteristics X_{t-1}.

This can be important in impact evaluation. Surveys for impact evaluations are often longitudinal in nature because many impact models are longitudinal in nature (such as the entire class of “within” causal models). It is important to remember, however, what adjustments like this strive to do in longitudinal settings. Cross-sectional surveys are designed to yield samples representative of some population at the time of the interview. Specifically, they are supposed to yield samples representative of the population that the selected units of observation represented at the time of interview. Non-response adjustments in this setting strive to bend representativeness back toward whatever population the sampling units selected represented.

Longitudinal surveys are designed to be representative of some population at the time of *baseline* (or first) interview. But the adjustments

$$\frac{n_t^R}{n_{t-1}^s}$$

bend representativeness back toward the population represented by the respondents as of the last round of interviews. These adjustments thus do not necessarily bend representativeness back toward the population represented by those *selected* at baseline, except in the case where there was no non-response at baseline (in which case there was no difference between the selected and respondent “populations” at baseline).

The sort of weight class analysis described thus far is based on independent calculations of non-response for each “cell” (or combination of values from X). In this it is somewhat akin to the calculation challenge of classical matching. Like classical matching, it can suffer from a curse of dimensionality. Table 2.10 shows how rapidly the number of cells can grow as the number of characteristics (and the categorical detail provided about each characteristic) grows. As we can see, just a few characteristics have led to 15,000 cells for which non-response might need to be calculated.

Table 2.10. Weight class adjustment: A few characteristics		
Characteristic	Categories	Cumulative Cells
Gender	2	2
Wealth quintile	5	10
Education level	6	60
Language spoken at home	5	300
Employment status	2	600
Age (in bands)	5	3000
Religion	5	15000
Total number of "cells"		15000

Suppose, however, that the sample at $t-1$ contained only 20,000 observations (we say "only," but that would be a huge sample in many fields). Assuming each of the cells in Table 2.10 are observed at time $t-1$ and definable at time t (i.e., there is at least one respondent at time t in each cell; more on this at the end of the paragraph), we would in principle have just 1.33 observations per cell. This can be a real challenge since the tiny numbers of observations in each cell might lead to wild estimates of cell response probabilities. This will in term lead to wild variation in the weights (remember, this is a non-response adjustment to the weights), and we have seen the consequences of weight variation. This would likely not have been the case had there been hundreds of observations in each cell. Moreover, the weight adjustment would be undefined for any cell for which

$$\frac{n_t^R}{n_{t-1}^s} = 0$$

This discussion highlights a tradeoff to weight class adjustment: the fewer characteristics one considers, the less convincing the adjustment and the more characteristics considered the more unwieldy the process becomes and the more operational and statistical complications it can introduce.

One popular solution to this (indeed, the preferred option for the authors) is to move toward a regression-based approach to non-response. At the core of this is the latent variable regression model

$$n_{i,t}^{R*} = \beta_0 + \beta_1 \cdot x_{1,t-1} + \beta_2 \cdot x_{2,t-1} + ... + +\beta_k \cdot x_{k,t-1} + \epsilon_t$$

where i is the i^{th} unit of observation interviewed at $t-1$ and $\{x_{1,t-1}, x_{2,t-1}, ..., x_{k,t-1}\}$ are the set of observed characteristics. $n_{i,t}^R$ equals 1 if the i^{th} unit of observation responds at time t and equals 0 otherwise.

This model is typically estimated via some binary regression model (for example logit, in which case we assume that ϵ_t follows a Type-I extreme value distribution, or probit, in which case we assume that ϵ_t is normally distributed) where the binary response variable $n_{i,t}^R$ is the dependent variable. The fitted (estimated) model is then used to predict the probability of response for each time t respondent given their characteristics $\{x_{1,t-1}, x_{2,t-1}, ..., x_{k,t-1}\}$. The inverse of this is the adjustment to their weight.[37]

[37] Though this captures the essence of the process, there can be more to it than this and more technical considerations

2.3.2 Post-Stratification Adjustment

So far, we have considered nonresponse adjustments, which address potential differences between the population represented by those selected and the population represented by those actually interviewed. Assuming that selection from the frame was done "correctly" (for example, in accordance with the requirements of probability sampling) then the population represented by the frame and by the sample are one and the same (abstracting away from any non-response and any need to make adjustments based on unequal probabilities of selection from the frame). The idea is that the frame covers the population of interest and hence those selected from it are representative of that population of interest.

But what if the frame did not wholly capture the population of interest? In particular, what if some members of the population of interest did not appear on the frame? This is the problem of frame under-coverage. Against the ultimate goal (estimates representative of the population of interest) this is a kind of failure of probability sampling: There are some members of the population of interest with no probability of selection because they don't appear on the frame from which selection is performed.

Sometimes, there is an auxiliary[38] data source (such as a census of the population of interest, a sample that is believed to be representative of the sample of interest, etc.) that provides information about the true distribution of types of individuals across the population of interest. Obviously, we can only observe the true distribution of types across the population of interest in terms of characteristics of its members that are observable and observed. Once again, let's call these X.

Assuming that these characteristics are also observed in the sample, one response to possible frame under-coverage has been to exploit auxiliary information about the distribution of characteristics X across the population of interest to address the consequences of that under-coverage. Crudely, the basic logic is to adjust weights to boost the information contribution to estimates of those under-represented in the sample compared with their actual presence in the population of interest. For the q^{th} type (in terms of a combination of characteristics from X), the adjustment would be related to

$$\frac{p_q^{POP}}{p_q^{SAMP}}$$

where p_q^{SAMP} is the proportion of the sample from the q^{th} type and p_q^{POP} is the proportion of the population of interest from the q^{th} type. This type of adjustment is called a post-stratification adjustment.

can be introduced. The interested reader will not have a hard time finding references for further discussion and examples of selection on observables-based weight adjustments. For some quick, easy (and web-available!) starting points, see perhaps Iannacchione, 2003; Iannacchione, Milne, & Folsom, 1991; Grau, Potter, Williams, & Diaz-Tena, 2006; Potter, Garu, Williams, Diaz-Tena, & Carlson, 2006. Though the authors admit that they have not kept close track of this literature in recent years, it would seem that this area is not one where settled best practice has been established around any number of technical considerations. Our own casual experimentation over the years tends to suggest in some gross sense that less rides on the particulars than the voluminous literature might suggest, but as we said, these were just casual experiments.

[38] In this instance, by "auxiliary" we mean a data source separate from our sample. In other statistical settings the word can carry other meanings.

Post-stratification is very common in population surveys in wealthier countries but less so in those from poorer and middle-income ones. The reason is simple: Auxiliary information sources are less readily available in the latter. However, as information generation processes like national census grow in quality and timeliness globally, post-stratification adjustments may become more widespread.

It is less clear that that will be the case with surveys for impact evaluations. Even in societies such as the United States, post-stratification adjustments are, to make a broad generalization, not that common in surveys designed to support impact evaluations. The reason is that in many cases, the population of interest (such as the participant population) are subnational groups for which there is no reliable auxiliary information on which to base a post-stratification adjustment.

To round out the discussion of weight adjustments to influence representativeness, a graphical summary might be in order. For inspiration we turn to a classic graphical device that the authors associate with Paul Biemer (we have seen it in several of his presentations; see, for instance Biemer & Christ, 2008). In Figure 2.13 we consider how weights can take us from the sample to the target population for a survey. First, a sample might not reflect the respondent population if the sampling

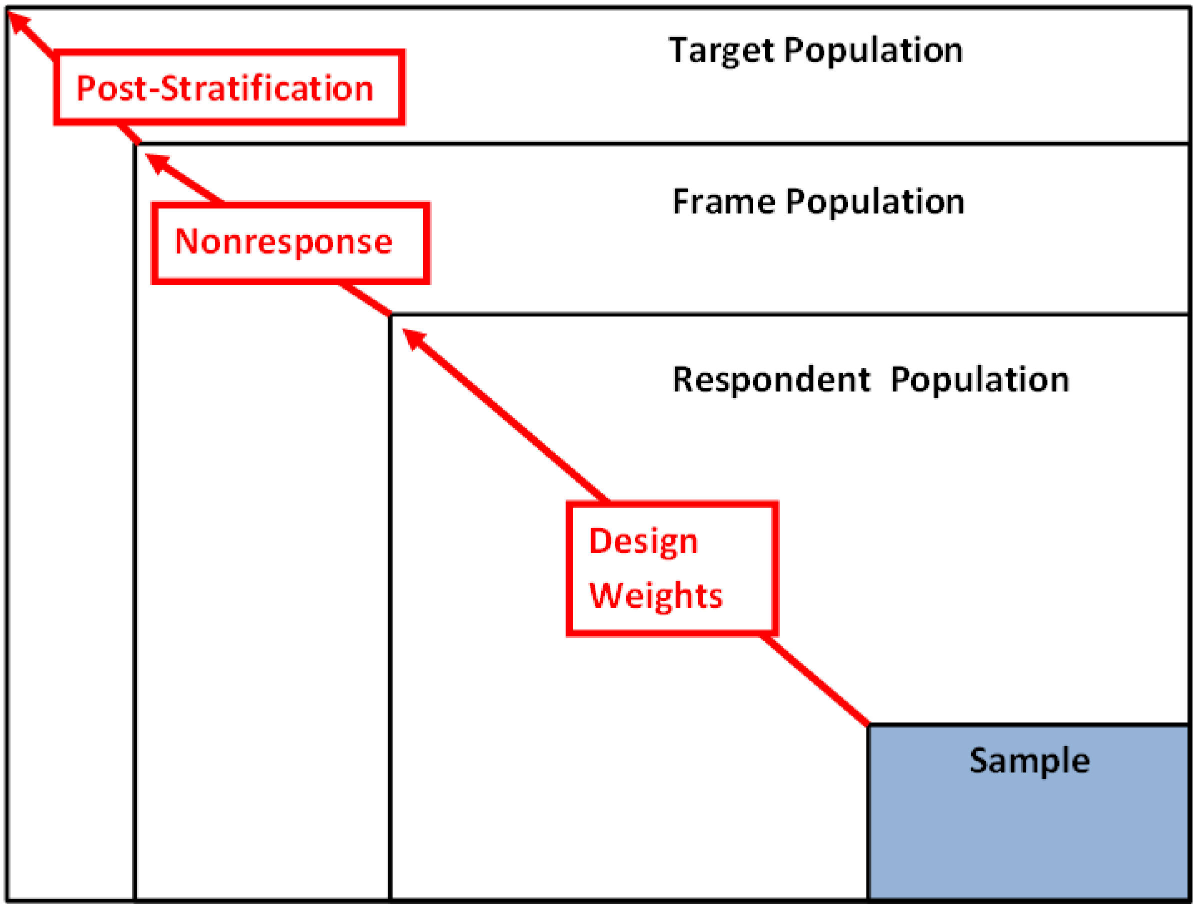


Figure 2.13. The role of different types of weights

units selected into the sample had different probabilities of selection. Design weights (which are the inverse of the design probability of selection) will make the sample representative of the respondent population. Some units selected from the frame will not respond, meaning that the respondent population might differ from the frame population. This gap is usually addressed with a non-response adjustment. Finally, due to frame under-coverage (whereby some members of the target population are not present in the frame to begin with) there can be a difference between the frame and target populations. It is this gap that post-stratification seeks to address.

2.3.3 Weight "Normalization"

We now turn to a topic that has, in our professional experience, been the source of massive confusion for those who are not familiar with sampling but must form estimates with samples that contain design-based weights. Let us begin with reminding ourselves once again what a design weight is. A design weight is the inverse of the probability of selection from the frame:

$$w = \frac{1}{p}$$

where p is the "design" probability of selection.[39] The weight, so conceived, has a behavioral interpretation: It is an estimate of the number of units in the population captured by the frame that a particular selected unit of observations represents. As such, this is sometimes called the "raw" weight, because it is still in terms of the original raw interpretation of a probability weight.

Consider a hypothetical example. After selection from a frame of all Americans (this could involve multistage selection), a design weight of 1500 for one American ultimately selected for the sample (call the person American 1) means that that selected American represents (or, more precisely, is estimated to represent) 1500 Americans. More to the point, that American represents around 1500 of his or her type of American. Suppose that another selected American (American 2) represents around 3,000 Americans (per their design weight). Together, the two Americans represent 4,500 Americans (not necessarily the same "type," however). Their relative weights are

$$\frac{w_2}{w_1} = \frac{3000}{1500} = 2$$

In other words, American 2 represents twice as many Americans as American 1, and hence will make twice the contribution to the weighted estimate

$$\frac{\sum_{i=1}^{2} w_i \cdot Y_i}{\sum_{i=1}^{2} w_i} = \frac{1500 \cdot Y_1 + 3000 \cdot Y_2}{4500} = \frac{1500 \cdot Y_1}{4500} + \frac{3000 \cdot Y_2}{4500} = \frac{1}{3} \cdot Y_1 + \frac{2}{3} Y_2$$

The "weighted count" (i.e., weighted number of observations) behind this estimation would be 4500.

[39] We focus on design weights not because non-response or post-stratification adjustments shouldn't be added before normalizing weights (they should be added before any normalization) but also because including them would make the conversation more cumbersome for no additional insight.

Suppose that, for some reason, it was felt that results would be more understandable if the weighted observations equalled the unweighted observations. To achieve this, we would normalize the weights where the normalized weight is

$$w_i^N = w_i \cdot \frac{n}{\sum_{q=1}^{n} w_q}$$

In the example we have been discussing n would equal 2. In fact, the normalized weight (so that that the weighted number of observations equals the total number of observations) for this example would be

$$w_1^N = 1500 \cdot \frac{2}{4500} = \frac{3000}{4500} = \frac{2}{3}$$

and

$$w_2^N = 3000 \cdot \frac{2}{4500} = \frac{6000}{4500} = \frac{4}{3}$$

So the actual process of normalization is fairly simple. Having completed it, the weighted number of observations is now

$$w_1^N + w_2^N = \frac{2}{3} + \frac{4}{3} = 2$$

Thus, the normalized weights satisfy the objective — the weighted number of observations now equals the actual number of observations.

As practitioners, we have found that normalization is a particularly misunderstood area of weighting with numerous popular (well, as popular as anything can be among researchers) myths surrounding it. With just this simple example of Americans 1 and 2, we can dispense with many of them.

- **Myth 1: Normalization changes the estimate of an average or proportion.** Normalization does not change estimates of averages or proportions. Consider our simple example. The weighted estimate using our newly minted normalized weights would be

  $$\frac{\sum_{i=1}^{2} w_i^N \cdot Y_i}{\sum_{i=1}^{2} w_i^N} = \frac{\frac{2}{3} \cdot Y_1 + \frac{4}{3} \cdot Y_2}{2} = \frac{2}{6} \cdot Y_1 + \frac{4}{6} \cdot Y_2 = \frac{1}{3} \cdot Y_1 + \frac{2}{3} \cdot Y_2$$

  The normalization has made no difference to the estimate of this average — it reduces to exactly the same thing as in the case where we used the raw design weight.

- **Myth 2: Normalization changes the number of observations used in estimation.** It should be self-evident that this is not the case based on our last derivation. In general, the weight estimate (regardless of what riff on the original design or raw weight you use) is

  $$\frac{\sum_{i=1}^{n} w_i \cdot Y_i}{\sum_{i=1}^{n} w_i}$$

  The number of observations used is n, which does not change with the weight w applied.

- **Myth 3: Normalization changes the efficiency of the estimator.** From what we derive in response to Myth 1, it should be clear that this is untrue. However, to further convince the reader, we offer a variation of Departure 4 from earlier in this chapter. Recall that in the original Departure 4, we found a design effect of 2.8176274 (computed from the variance estimates across the 1,000 simulated sample). We then offer the following twist: we redo Departure 4, except that on each simulated selection we normalize the design weight so that the weighted observations across the sample is 250, the same as the actual number of observations in the sample. The result is...the same design effect. The reason for this is that what matters is relative variation in weight. The relative variation of the weights is unchanged by normalization. Beginning with the basic normalization formula

$$w_i^N = w_i \cdot \frac{n}{\sum_{q=1}^{n} w_q}$$

consider the relative weights of observations i and j:

$$\frac{w_i^N}{w_j^N} = \frac{w_i \cdot \frac{n}{\sum_{q=1}^{n} w_q}}{w_j \cdot \frac{n}{\sum_{q=1}^{n} w_q}} = \frac{w_i}{w_j}$$

In other words, the relative normalized weights are the same as the relative design weights. A crude but surprisingly useful rough indication of the variance consequences of weights (particularly less informative weights) is the ratio of the largest to the smallest weight.

- **Myth 4: Normalization has a behavioral interpretation.** A raw or design weight has a behavioral interpretation (it is an estimate of the number of members of the frame population that a given selected unit represents). While the relative weights always have behavioral meaning, the magnitude of a normalized weight has no such meaning. In fact, for a given sample or subsample, we can normalize so that the weighted number of observations equals any positive value that we want. In other words, we have thus far focused on normalizing so that the weighted number of observations equals the actual number of observations, leading to the normalization formula

$$w_i^N = w_i \cdot \frac{n}{\sum_{q=1}^{n} w_q}$$

However, normalization can generate whatever weighted number of observations we want. The more general formula is

$$w_i^N = w_i \cdot \frac{\Omega}{\sum_{i=q}^{n} w_q}$$

where we can generate whatever weighted count we want by setting Ω appropriately.

- **Myth 5: Normalization is always harmless or without consequence.** Normalization can sow confusion and, when only normalized weights are available, limits what can be successfully estimated. It should be obvious that, to the extent that normalization occurs across many subsamples, one can end up with many weights in the sample, and a lot of confusion about which weight should be used when. We have seen that normalization makes no differ-

> ence to a weighted average or proportion. But what if our goal was to estimate a population total? The weighted estimated would then be
>
> $$\sum_{i=1}^{N} w_i \cdot Y_i$$
>
> It should be obvious from our work on the simple example that the normalized weight does not necessarily yield the population total. For estimating a total, the clear behavioral interpretation of a design weight is actually really important.

We would like to say that these myths are merely comical, but unfortunately they are sometimes substantive in that they can influence the decision of whether and how to normalize weights.

So why would one normalize weights? The original motivation for weight normalization was related to the technical limitations of early statistical software (Heeringa, West, & Berglund, 2010). In early statistical software capable of incorporating weights into estimation (for example, of average with samples), weights were handled in the analysis as if they were frequency weights rather than probability weights.[40] In analyses with frequency weights, each observation is duplicated as many times as its frequency weight; for instance, an observation with a frequency weight of 1,000 will be treated for estimation purposes as if it is actually 1,000 observations. This practice, however, does not affect the estimation of survey variables.

Recall that the unbiased estimator of the population average of a variable Y with probability weights is given as

$$\frac{\sum_{i=1}^{n} w_i \cdot Y_i}{\sum_{i=1}^{n} w_i}$$

where Y_i is the observed value of Y for i^{th} individual, and w_i is the probability weight of i^{th} individual, for $i = 1, ..., n$. Now suppose that for estimation of the population mean the probability weights are actually treated as frequency weights such that each observation is duplicated w_i times in the dataset. Then the population average of Y is estimated as

$$\frac{\sum_{i=1}^{n} \sum_{i=1}^{w_i} Y_i}{\sum_{i=1}^{n} \sum_{i=1}^{w_i} 1} = \frac{\sum_{i=1}^{n} w_i \cdot Y_i}{\sum_{i=1}^{n} w_i}$$

which is exactly the same as the unbiased estimator based on probability weights. Therefore, the two estimators are both unbiased and yield the same estimate for the population average of Y.

A problem arises when we estimate the sampling variance. Recall that the sampling variance is

[40] A frequency weight represents the number of members of the sample a given observation included in the sample represents. For instance, in a survey with only three recorded characteristics, age, a poverty indicator and gender, it might have been the case that 13 poor women ages 37 were interviewed. Particularly in an era where computer memory was dear, it might have been the practice to actually leave only one of these women in the sample with a frequency weight of 13, indicating that this observation represents 13 identical observations in the sample. As such, frequency weights generally took on only positive integer values. Frequency weights are also generally not estimates. Frequency weights are distinct from the sampling or probability (of selection) weights that we have considered thus far, which are an estimate of the number of individuals in the population that each observation represents.

estimated as

$$var\left(\hat{\bar{Y}}\right) = se^2\left(\bar{Y}\right) = \frac{s^2}{n}$$

where $\hat{\bar{Y}}$ is our estimate of the population mean of Y. When the probability weights are treated as frequency weights, this becomes

$$var\left(\hat{\bar{Y}}\right) = se^2\left(\hat{\bar{Y}}\right) = \frac{s^2}{\sum_{i=1}^{n}\sum_{i=1}^{w_i} 1} = \frac{s^2}{\sum_{i=1}^{n} w_i}$$

Note that the denominator has become the sum of weights ($\sum_{i=1}^{n} w_i$) instead of the sample size (n).

Now let's look at the implication of the change in the denominator between normalized and un-normalized weights ($w_i^{normalized}$ and $w_i^{un-normalized}$, respectively). Under un-normalized weights the denominator becomes

$$\sum_{i=1}^{n} w_i^{un-normalized} = \sum_{i=1}^{n} \frac{1}{p_i} = N > n$$

Therefore the sampling variance is underestimated when un-normalized weights are applied as frequency weights. This problem in frequency weights was then addressed by introducing normalization to the weights. Under normalized weights the denominator becomes

$$\sum_{i=1}^{n} w_i^{normalized} = \sum_{i=1}^{n}\left(w_i^{un-normalized}\right)\cdot\left\{\frac{n}{\sum_{i=1}^{n}\left(w_i^{un-normalized}\right)}\right\} = N \cdot \frac{n}{N} = n$$

which scales back the sum of the weights to the sample size, the correct denominator.

This problem in early statistical software is no longer an issue with modern commercial statistical software packages. The newer packages actually generally allow users to specify how the weights should be treated. Nonetheless, the practice of using normalized weights has for some reason long outlived the original reason for doing so. In some sense, normalization of weights is thus an excellent example of the "QWERTY" phenomenon.[41]

Weights can be normalized across the sample or across any subsample of that sample. In general, a weight (design or normalized) can generate statistics representative of the population captured by any subsample of any sample or subsample across which weights have been normalized, but not across populations for which weights have been separately normalized.

While precise that sentence while precise can be pretty confusing, so we offer a simple example. Suppose that we have a sample S, and that there are two subsamples of interest, A and B. Weights

[41] Virtually every keyboard on earth has the string "QWERTY" in its top line of characters. The reasons for this extend far back into the 19th century, when typewriters were first developed. In those days, pressing the keyboard prompted an arm with the desired character to swing out and strike the page through an ink ribbon. Unfortunately, in those early typewriters the arms were made of steel that was either too soft or brittle, leading to damage when two arms struck each other due to typing too quickly. This led to the development of keyboards that slowed typing enough to avoid such collisions and, as it turned out, the keyboard arrangement with QWERTY in the top row slowed typing sufficiently. The motivation for QWERTY has thus long passed, and yet the top row of the keyboard of the 21^{st} century laptop on which the authors write this footnote still reads "QWERTY."

(normalized or design) at the level of sample S can produce estimates of averages or proportions representative of the populations represented by subsamples A or B: simply compute the weighted average across subsamples A and B separately, but using the overall sample-wide weights. Similarly, estimates of averages or proportions representative of the populations representative of subsamples of, say, subsample A can be computed with weights normalized across subsample A so that, for example, the weighted number of observations equals the actual number of observations in A. What one cannot do is to somehow combine weights separately normalized for subsamples A and B. Instead one would need to normalize a weight across A and B jointly.

Or one could eschew normalization altogether. That might spare users of the data a lot of confusion.

2.3.4 Weight "Trimming"

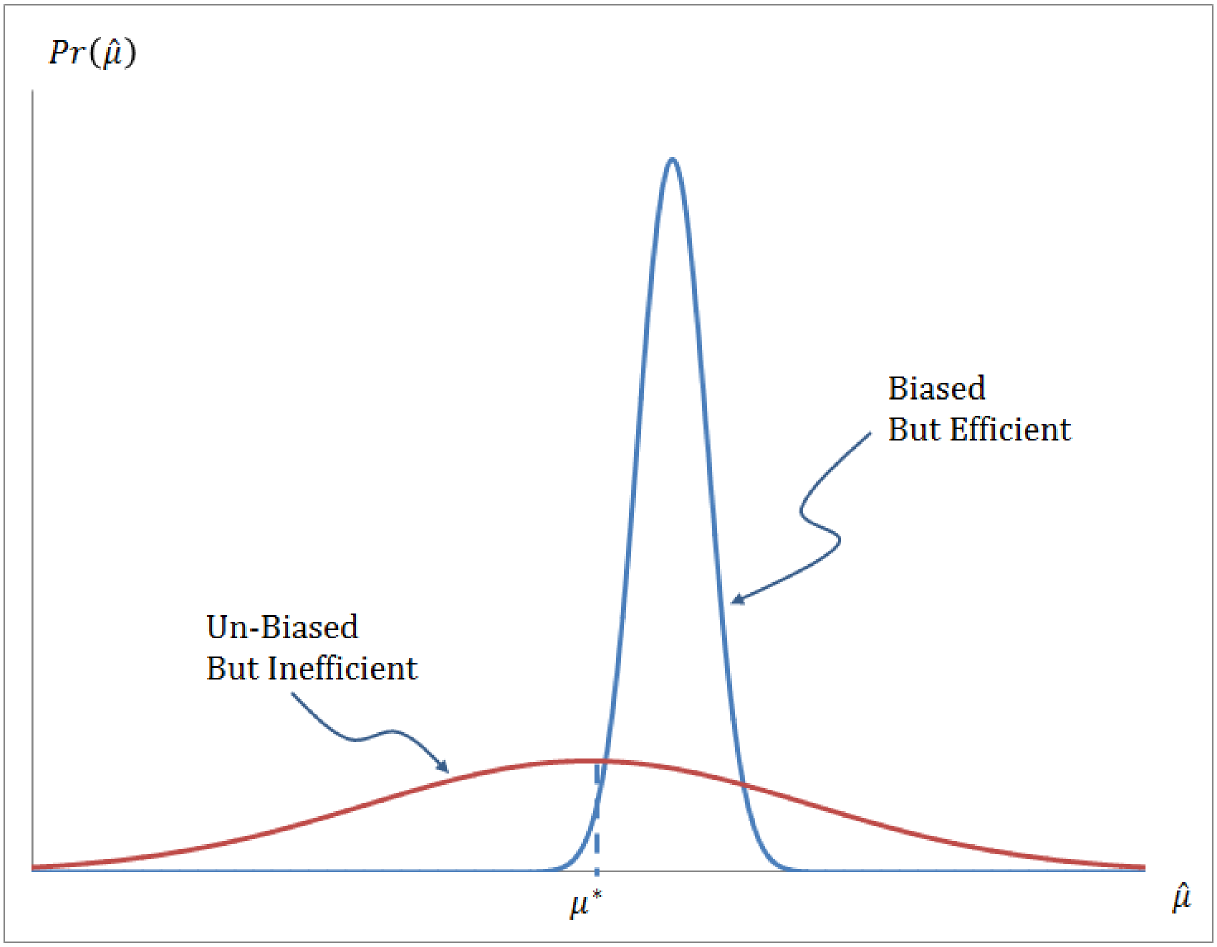


Figure 2.14. Bias versus efficiency

We have seen that unequal probabilities of selection across those selected for a sample gives rise to weights that also vary across them. We have also seen that this has consequences in terms of elevated sampling variation to estimates. Some have suggested that the solution should be to "trim" extreme weight values back toward the mean for all weights.

The problem with this strategy is that, assuming probability sampling has been followed and probabilities of selection have been correctly modelled for crafting weights, the design weights are the statistically correct weights. Perhaps one could justify this in cases where the wild variation is driven by, for instance, huge adjustments to weights due to small cell estimates of response probabilities under a weight class adjustment scheme, but in general, for all of the sophistication of the discussion of trimming, trimming tends to be ad hoc and atheoretical. The real logic is simply to get rid of wild weights.

Which is not to say that weight trimming is altogether unjustified. One way of thinking about weight trimming is in terms of the old "bias versus efficiency" tradeoff. Sometimes there is a biased estimator that is much more efficient (i.e., lower sampling variation) than the most efficient available unbiased estimator.[42] This possibility is illustrated in Figure 2.14. The biased estimator might be, well, biased, but in practice it will often generate estimates much closer to the true population value μ^* than the unbiased alternative. This should manifest itself in a mean squared error

$$E\left((\hat{\mu} - \mu^*)^2\right)$$

that is actually lower for the biased estimator.[43]

We cannot offer concrete recommendations in terms of what to do in any given sample. The sampler must use his or her judgment given the estimator to be used, degree of weight variation, etc. in deciding whether weight trimming is justified. For what it is worth, the authors have never trimmed a weight.

2.3.5 Multiple and Overlapping Frames

One complication we wish to mention is selection from multiple, overlapping frames. By "overlapping" we mean that a sampling unit can be selected from more than one frame. A few examples are:

- In political polling, consumer sentiment polling, and other opinion polling it is common to rely on telephone surveys. These often involve random selection of phone numbers from frames of landline and mobile phone numbers. Some individuals in the population have only one type of phone, but for those with more than one type of phone their are multiple ways that they can be selected (combining the frames does not alleviate this problem unless we can tell the numbers common to a single individual).

- There is sometimes a temptation to compensate for some failure to selected sampling units by selecting more. Recently, one of the authors was involved in an impact evaluation survey

[42] Indeed, there is an entire tradition of testing in impact evaluation usually referred to collectively as "Hausman-type" tests that are, at root, based on comparing estimates from an unbiased (or consistent) but inefficient estimator with a (potentially) biased (or inconsistent) but efficient estimator.

[43] The mean squared error is a good way of answering the question of how much the estimates from this estimator typically differ from the true population value, regardless of central tendency.

in which it was suggested that selected households that were "too hard" to contact would be replaced with new selections from the household list. The problem with this is that there would then be two ways that each household could have been selected: during the original selection and during the replacement round of selection. Since selection is done twice over the same frame, the overlapping frame issue emerges.[44]

- One of the authors was once involved in the design of a survey for an impact evaluation where the following selection mechanism was proposed. From a list of the areas (in this case, villages) where the program would be operating, a sample of primary sampling units for populations exposed to the program would be selected. To select primary sampling units for areas not exposed to the program (i.e., primary sampling units for control populations) it was proposed that non-program areas adjacent to each selected program area be listed and, for each program area, one would be selected. This is the situation illustrated in Figure 2.15. The program areas are indicated by the blue rectangles, while adjacent non-program areas are indicated by the yellow ones labeled "NP." The challenge arises with non-program areas that are adjacent to more than one project area. This is illustrated by the non-program area with the dashed red border in Figure 2.15. This particular non-program area could be chosen through either of the program areas shown, and hence in two different ways.

- Suppose that one wished to sample households influenced by a program by selecting from

[44] Another problem with this suggestion is that it is verging on introducing an element of convenience sampling to the selection procedure.

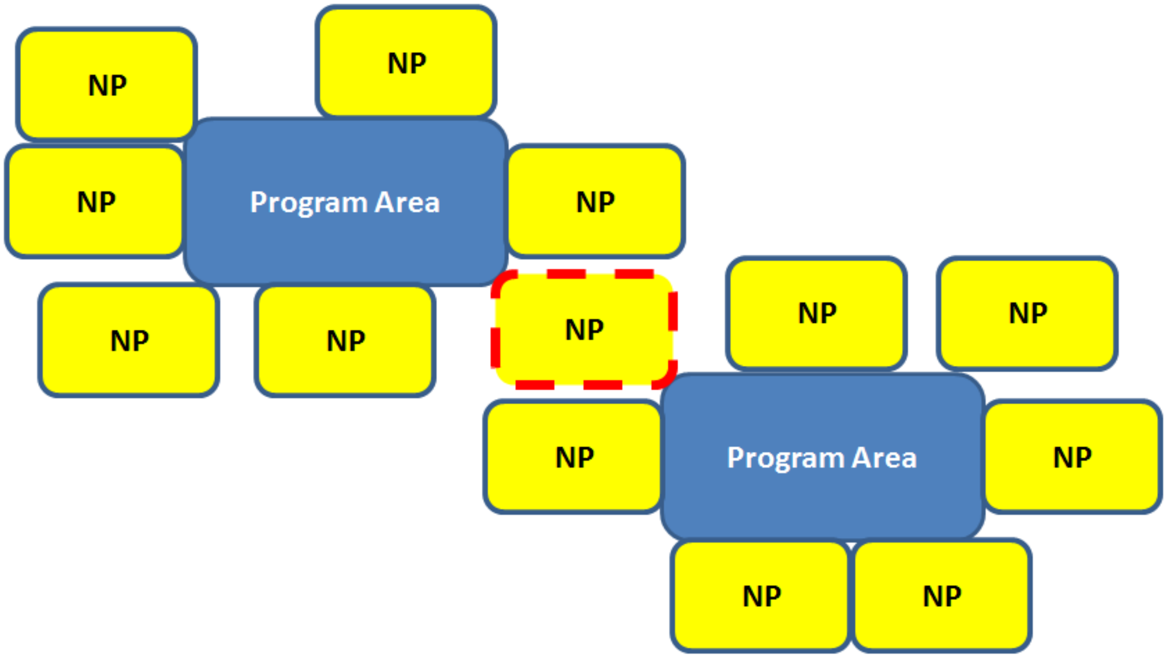


Figure 2.15. An instance of dual frames

a list of individuals enrolled in the program and then interviewing all individuals in any selected participant's household. Suppose, however, that there was potentially more than one participant per household. This means that there is more than one way that the members of households could be selected.

As these examples make clear, the overlapping frame problem is sometimes obvious, but in other cases can be less clear. One must always consider the possibility of it when considering a proposed selection design. When we evaluate a proposed selection design, among the questions we ask ourselves as a matter of standard practice is:

Can some or all of the sampling units be selected in more than one way?

This is not a question that merits a snap answer but instead careful reflection, because the mechanism by which this might occur might not be obvious at first glance.

To help to wrap one's mind across this problem, we offer a simple example motivated by a problem that one of the authors encountered in an actual impact evaluation problem. Suppose you wanted to estimate total income Y for some population represented by two frames (meaning that, between them, every unit in the population fell into at least one of the two frames). Call these frames A and B.

Assume that you choose n_A units from frame A (out of a total of N_A units in frame A) by epsem sampling. The probability of selection is thus

$$\frac{n_A}{N_A}$$

and the weight for selected units would be

$$\frac{N_A}{n_A}$$

If y_A is the average income for the units in frame A, then an unbiased estimate of total income Y_A for the population represented by A would be

$$Y_A = (n_A) \cdot \left(\frac{N_A}{n_A}\right) \cdot y_A = N_A \cdot y_A$$

By exactly the same logic, the estimate of the total income Y_B for the population represented by frame B would be estimated in an unbiased fashion by

$$Y_B = (n_B) \cdot \left(\frac{N_B}{n_B}\right) \cdot y_B = N_B \cdot y_B$$

If frames A and B did not overlap (i.e., no one appeared in both frames) but together included all members of some population of interest, then a valid estimate of total income for that population would be

$$Y = N_A \cdot y_A + N_B \cdot y_B$$

So far this seems reasonably straightforward.

Let us suppose instead, however, that A and B overlap. Specifically, assume N_{AB} individuals are in both frames, N_{AA} individuals are just in A and N_{BB} individuals as just in B and the average incomes are: Y_{AA} for those only in frame A; Y_{AB} for those in frames A and B and Y_{BB} for those only in frame B.

Ideally we would independently sample from three frames (the frame for those just in A, the frame for those just in B, and the frame for those in both A and B). Applying exactly the logic above, an unbiased estimated of Y based on sampling from the three frames would be

$$Y = N_{AA} \cdot y_{AA} + N_{BB} \cdot y_{BB} + N_{AB} \cdot y_{AB}$$

However, suppose that we naively estimated from just A and B without recognizing the overlap. Then we have

$$Y_A = (n_A) \cdot \left(\frac{N_A}{n_A}\right) \cdot y_A = N_A \cdot y_A = N_{AA} \cdot y_{AA} + N_{AB} \cdot y_{AB}$$

and similarly

$$Y_B = (n_B) \cdot \left(\frac{N_B}{n_B}\right) \cdot y_B = N_B \cdot y_B = N_{BB} * y_{BB} + N_{AB} \cdot y_{AB}$$

An estimate of the total income from such a nave estimation would be

$$Y_A + Y_B = N_{AA} \cdot y_{AA} + N_{AB} \cdot y_{AB} + N_{BB} \cdot y_{BB} + N_{AB} \cdot y_{AB} = N_{AA} \cdot y_{AA} + 2 \cdot N_{AB} \cdot y_{AB} + N_{BB} \cdot y_{BB}$$

$$> N_{AA} \cdot y_{AA} + N_{BB} \cdot y_{BB} + N_{AB} \cdot y_{AB} = Y$$

We thus overestimate total income when we do not adjust weights for the overlapping units. Intuitively, we are not correctly accounting for the fact that, other things being equal, those in both frames have an exceptionally high probability of selection.

In many real world settings, dealing with this problem is potentially tedious. The general solution is that the weights for sampling units subject to overlapping frame selection (that is, that could have been selected out of more than one frame) is a convex function of the weights that they would have based on their selection probability from each frame in which they appear (i.e., their weight from each frame as if they had only been selected in that frame). Strictly speaking, the nature of the combination can be a complex one. For instance, the convex combination can depend on an inverse function of the variance of the variable of interest (i.e., the variable behind the population parameter of interest). In a survey with many such variables of interest, this could rapidly require a huge number of weights. That said, some (such as Skinner & Rao, 1996) have proposed near optimal or approximate approaches that yield one weight common to all outcomes.[45] The details of these various possibilities are probably too much of a digression for the present purposes.

For the present purposes, we simply wish to make the reader aware of this possibility and to be on the lookout to recognize it (as we have seen, it can be subtle!). Overlapping frame circumstances are to be avoided in the sense that they offer potential complication for little gain. Therefore, given the choice between two alternative selection schemes, one of which involves an overlapping frames

[45] The authors thank Diana Stukel for a stimulating exchange regarding this subject.

circumstance but is otherwise appear equally compelling (for example, one does not have a clear efficiency advantage over the other), we would tend to avoid the scheme involving overlapping frame selection.

2.4 Spillover

In this section we discuss a potential complication that, in our experience in the international health and family planning context, is attracting increasing attention in the design phase of impact evaluations. Specifically, one concern is that the participation status of one observational unit can influence the potential outcomes $\{Y^1, Y^0\}$ of another observational unit.

This possibility is usually referred to as spillover, though we have sometimes heard other terms applied to it (including contagion, externalities, etc.). Our initial statement of this possibility was in the broadest terms possible. We expressed it in terms of a potential impact from *any* unit or agent's participation status on either of *any* other unit's potential outcomes (i.e., Y^0 or Y^1). This allows for a lot of possibilities. For instance, the outcome under participation (Y^1) for one individual could change if another were to participate. At the philosophical level (at least) the idea that one person's potential outcomes are influenced by another's participation could evoke all kinds of conceptually interesting hypothetical possibilities.

In practice, however, the concern, and hence discussion around it in the design phase of impact evaluations, is typically narrower. Usually the concern is that the participation behavior, circumstances or experiences of participants will have some kind of impact on the outcome in the absence of participation, Y^0, of non-participants. Moreover, there is usually less concern with the impact of any one particular individual participant's behavior, circumstances, or experiences on any particular non-participant i's observed outcome

$$Y_i = P_i \cdot Y_i^1 + (1 - P_i) \cdot Y_i^0 = 0 \cdot Y_i^1 + (1 - 0) \cdot Y_i^0 = Y_i^0$$

Instead, the general notional sense is that the collective behavior, circumstances, or experiences of participants as a group will impact the observed outcomes of non-participants as a group, or perhaps some subset of them.

This has been a vague enough exposition that perhaps it is worthwhile to suggest some concrete examples. A few highly plausible possibilities might include the following:

- An immunization program is likely to influence the health of those not immunized by generally reducing the transmission of the targeted diseases.

- Similarly, virtually any health program that targets a channel of health subject to externalities (whereby the health of one person can depend on the behavior, circumstances, or outcomes of another) might influence the health of nonparticipants (examples might include a sanitation program that improves the health of those who already enjoyed modern sanitation by reducing the transmission of disease, a smoking cessation program that reduces the exposure to second-hand smoke of non-participants, etc.).

- A successful (in the sense of reducing crime) policing program might actually worsen crime in adjoining communities not exposed to the program as criminals shift to the more favorable circumstances they provide.

- More generally, programs that influence the behavior of third party agents whose own actions might in turn influence potential outcomes can generate spillover as those agents react to the program.

- A labor market training program can influence those who do not participate by, for instance, increasing the supply of trained labor with whom they must compete in the labor market.

- Similarly, virtually any program that can influence general economic or market circumstances, even on a local scale, can influence the welfare of non-participants (examples could include a cash transfer program that might influence the incomes and other socioeconomic outcomes of non-participant households through a general equilibrium effect as those with additional income spend more in their communities, agricultural value-chain programs that shift the supply curve for the products involved and hence the market conditions of non-participant farmers, a health insurance scheme that shifts the demand curve for health care and hence the health care prices confronting those not participating in the program, etc.).

- A behavior change and communication program focused on family planning in urban areas of a lower income nation might influence family planning use through the diffusion of resulting norms (from the program) that are more favorable to use of modern contraception.

- Similarly, virtually any program designed to influence attitudes of participants can influence the outcomes experienced by non-participants through social transmission (examples might include a gender equity program that influences general cultural norms and hence the behavior of non-participant households, a healthy eating program that leads to wider social understanding of nutritionally sound eating practices, an agricultural extension or teacher training program that leads to the dissemination of better practices even to those who did not participate, etc.).

This should be viewed as a highly incomplete list of possibilities. The possibilities for spillover are limited only by the imagination. Indeed, even the programmatic examples raised in this list could involve many other channels for spillover, and a given program can generate spillovers along numerous channels (including health externalities, general equilibrium economic consequences, changing norms, attitudes or social conventions, the adaptive behaviors of third party agents, etc.)

Two truths of potential spillovers should be evident from these examples. First, some types of spillover might actually be intended. For instance, one motivation for the urban focus of some family planning programs in the last few years in lower-income nations (such as the Bill & Melinda Gates

Foundation-funded Urban Reproductive Health Initiative in select cities of Uttar Pradesh, India, Kenya, Nigeria, and Senegal) has been the hope that successful urban family planning programs might enjoy impact beyond the cities as spreading norms of usage of modern family planning diffused to surrounding rural areas. Below we will often approach the possibility of spillover as a kind of nuisance that can complicate estimations of the impact of direct participation in the program. That being said, to the extent that spillovers were a hoped-for feature of the program, capturing them might be necessary for understanding total impact and hence, for instance, the broadest (and thus perhaps the fairest) measure of program effectiveness for the purpose of assessing crucial decision point parameters such as relative cost-effectiveness.

Second, spillovers might be a good or bad thing for those non-participants experiencing them. When herd immunity is achieved with respect to, for instance, rabies in dogs, unvaccinated dogs clearly experience a positive spillover in the form of a much diminished probability of contracting rabies. On the other hand, communities experiencing a surge in crime as criminals exit nearby areas due to an effective new policing program or strategy are not experiencing a positive spillover. Indeed, the anti-vaccination movement in the United States can be viewed as a kind of perverse informal program[46] (where one "participates" by shunning vaccination) which has involved considerable spillovers as local outbreaks of vaccine-preventable diseases have occurred largely due to the declining vaccination rates. While these outbreaks have generally been concentrated among the unvaccinated, one must remember that the unvaccinated are not necessarily all participants in this "program" so defined (since some are unvaccinated not because they choose to shun vaccines but because they are too young to receive them or unable to be vaccinated for medical reasons) and, to the extent that many vaccines are not completely effective, even those who are vaccinated can be affected by this nasty spillover.

In the language of traditional causal estimation "theory" the idea that the participation decision of one individual (or decisions of several individuals) should not influence the potential outcomes of another person is usually referred to as the stable unit treatment value assumption (as in the program impact $Y_j^1 = Y_j^0$ one individual or "unit" j is stable in the face of the participation decisions of other units $i \neq j$) or, in short, SUTVA.[47] Violations of SUTVA undermine a basic assumption that is a departure point for many theoretical discussions of impact evaluation estimation theory.

Despite the importance of this topic (at least in the development phases of evaluations in which we have been involved) we have seen surprisingly little written about it in the peer-reviewed literature. The most comprehensive treatment we have seen (and which we highly recommend to

[46] Describing the anti-vaccination movement as a program might seem strange until one considers that it is promoted in large part by a kind of broad and essentially decentralized and activist-driven behavior change and communication agenda highlighting the supposed risks of vaccines.

[47] Formal statements of SUTVA often also include mention somehow of the idea that treatments are comparable between individuals. In other words, treatment should mean the same thing across individuals. This precludes the idea that, for instance, the potential treatment before two different individuals differs in nature. Where this assumption can get a bit fuzzy is the case of a dose-response type relationship. We generally take the view that this can be reasonably captured within the potential outcomes framework as long as the only thing that is truly varying is the dosage or treatment intensity for the same treatment. This can admittedly get a bit complicated if intensity is allowed to vary continuously; we usually think of discrete differences in dosage or intensity. However, not all statements of SUTVA that we have seen would seem to address this.

the reader) is Angelucci and Di Maro (2015). Although Angelucci & Di Maro seem particularly motivated by the possibility of spillover in the context of cash transfer examples, they offer many hypothetical examples under which spillover effects could emerge (some of which appear in our list of such possibilities). For the moment, we consider the issue from an angle that they pursue.

To begin with, Angelucci and Di Maro (2015) seem to root much of their conceptual approach to the problem from what seems a quite reasonable likelihood for spillovers in practice: the likelihood of spillover (or, in more formal terms, failure of the portion of SUTVA concerned with the possibility that one individual's potential outcomes can be influenced by the participation decision of another) increases with proximity. Proximity can mean a lot of things. For instance, in the case of diffusion effects along the lines pondered above spillover might involve social or cultural network proximity. Although that might at first glance seem a typically unwieldy social science construct, in the context of diffusion it could simply mean the suggestion that, for instance, spillover of attitudes, practices, and beliefs from urban to rural areas might be strongest or most plausible in those rural areas that are the setting for the most circular migration from cities, for example.

That being said, most of the time the focus is on geographic proximity, both in terms of our own experience in the design of impact evaluations and the narrative of Angelucci & di Maro (2015). Put simply, the closer in geographic space two agents are, the more likely, other things being equal, their individual behavioral choices are to influence each other. This becomes a particularly relevant concern when one considers that the data design behind many impact evaluations has a cluster element. It is often the case that in some clusters there are, per the definition of participation,[48] both participants and non-participants. Given that clusters are typically small geographically (more far-flung clusters would undercut one of the main advantages of the cluster design,[49] the opportunity for the spillover of program effects from participants to non-participants within the same cluster is natural and plausible.

Before exploring this possibility, we dispense with a few preliminaries. To simplify, let us assume that program participation is randomly assigned. Ignoring spillover for a second, this means that

$$E\left(Y^1|P=1\right)=E\left(Y^1|P=0\right)=E\left(Y^1\right)$$

and

$$E\left(Y^0|P=1\right)=E\left(Y^0|P=0\right)=E\left(Y^0\right)$$

By making this assumption we rule out the possibility of selectivity/confoundedness, or that program participation P is correlated with potential outcomes $\{Y^1, Y^0\}$.

Had we not made this assumption, it would be likely that

$$E\left(Y^0|P=1\right)\neq E\left(Y^0\right),$$

[48] Essentially, the focus here is on participation somehow conceptualized at a level of the individual, household or some unit below the cluster level. The present discussion is less motivated by the idea that all members of a cluster are treated as participants if the program operated at all in their cluster, a possibility to which we will return later in the discussion.

[49] Namely, lowered field costs as interviews and other field work can be conducted in smaller, more compact settings.

$$E\left(Y^0|P=0\right) \neq E\left(Y^0\right),$$
$$E\left(Y^0|P=1\right) \neq E\left(Y^0|P=0\right),$$
$$E\left(Y^1|P=1\right) \neq E\left(Y^1\right),$$
$$E\left(Y^1|P=0\right) \neq E\left(Y^1\right)$$

and

$$E\left(Y^1|P=1\right) \neq E\left(Y^1|P=0\right)$$

To summarize, if program participation was not random (meaning that participants and non-participants likely systematically differ in terms of their potential outcomes), then the observed experiences of participants would not necessarily be indicative of the outcomes under participation Y^1 that either non-participants or the general population would have experienced had they participated. In addition, the observed experiences of non-participants would not necessarily be indicative of the outcomes under non-participation Y^0 that either non-participants or the general population would have experienced had they not participated. This could complicate attempts to estimate many common impact parameters, such as the average treatment effect ($E(Y^1 - Y^0)$) or average effect of treatment on the treated ($E(Y^1 - Y^0|P = 1)$), using simple averages of the outcomes observed from participants and non-participants to estimate $E(Y^1)$, $E(Y^0)$, $E(Y^0|P = 1)$, etc.

Assuming that program participation is randomly assigned thus allows us to rule out such selection bias. In other words, program participation will not be confounded with potential outcomes. Thus, for instance, we could estimate the average treatment effect as the difference in the sample average observed outcomes from representative samples of participants and non-participants. This allows us to focus on potential complications from spillover such as bias to estimates of treatment effects without worrying that part of that bias is due to selectivity.

Define

$$S = \begin{cases} 1 & \text{if participation patterns in a cluster is such that spillover occurs} \\ 0 & \text{otherwise} \end{cases}$$

where we define spillover as some impact on Y^0 for non-participants as a result of the participation patterns in the cluster. For now we remain agnostic about what those patterns could be. That said, to help fix ideas, it could be that a certain percentage of the cluster must participate before spillover can occur. Alternatively, it could be that there is some absolute threshold of participation in terms of actual numbers participating before spillover can occur. The point is that there is some threshold of activity in terms of participation that must occur before spillover to Y^1 for non-participants can occur, and if $Z = 1$ that threshold has been met.

Our focus is on the possibility of spillover from participants to the observed outcome for non-participants (Y^0). We are thus simplifying the discussion by restricting the channels for spillover that we consider. For instance we ignore the possibility of some sort of synergies between participants that could serve either to amplify or attenuate Y^1 in the event that $S_j = 1$. Once this possibility is considered, one would also have to entertain the notion that the such spillovers could impact Y^1 for non-participants. While such possibilities are certainly conceptually interesting, we ignore them to focus on the possibility of spillover from the program to the observed outcome for non-participants. This is the channel for spillover that we have found most often to be the concern when designing impact evaluations. Nonetheless, our framework could still be applied to consideration of some alternative channels of spillover.

Having introduced this new spillover indicator, we can consider an elaboration on the expectation of the potential outcomes:

$$E\left(Y^1|S\right)$$

and

$$E\left(Y^0|S\right)$$

Since our focus is on spillover to observed outcomes for non-participants, we simply assume that

$$E\left(Y^1|S\right) = E\left(Y^1\right)$$

In other words, we rule out the possibility of spillover among participants (at least in terms of mean outcomes). Randomization of treatment then implies that

$$E\left(Y^0|S, P = 1\right) = E\left(Y^0|S, P = 0\right)$$

We thus dispense with the possibility that whatever problems we encounter from spillover will involve selectivity into participation status P.

Suppose that we wish to estimate the average treatment effect

$$E\left(Y^1 - Y^0\right) = E\left(Y^1\right) - E\left(Y^0\right)$$

where the right-hand side of the equality stems from the additive separability of expectations.[50] Suppose as well that we wish to estimate this in a cluster with spillover in effect ($S = 1$). We have already decided that

$$E\left(Y^1|S, P\right) = E\left(Y^1|S\right) = E\left(Y^1\right)$$

where the first equality holds due to the randomization of participation P and the second holds because we have decided to rule out the possibility of spillover to the potential outcome under participation. What all of this establishes is that we should be able to obtain an unbiased estimate of $E(Y^1)$ for the cluster with a simple sample mean of the observed outcomes ($Y = P \cdot Y^1 + (1-P) \cdot Y^0 = 1 \cdot Y^1 + (1-1) \cdot Y^0 = Y^1$) among participants.

[50] To refresh, if A and B are random variables, $E(A \pm B) = E(A) \pm E(B)$.

The problem comes when we try to form an estimate of

$$E(Y^0)$$

What we observe in the cluster is the outcome for non-participants given that they experienced spillover. We thus can estimate

$$E\left(Y^0|S=1,P\right)=E\left(Y^0|S=1\right)$$

We cannot go further than this as was the case when we approached estimation of $E(Y^1)$. If we estimate the outcome in the absence of participation using a simple mean of the observed outcomes ($Y=P\cdot Y^1+(1-P)\cdot Y^0=0\cdot Y^1+(1-0)\cdot Y^0=Y^0$) for non-participants in this cluster (and estimate $E(Y^1)$ with the sample mean of observed outcomes for participants in the cluster), the treatment effect we actually estimate is

$$E\left(Y^1-Y^0|S\right)=E\left(Y^1|S\right)-E\left(Y^0|S\right)=E\left(Y^1\right)-E\left(Y^0|S\right)$$

Assuming that the spillover effect

$$E\left(Y^0|S=1\right)-E\left(Y^0|S=0\right)\neq 0$$

we are not estimating the average treatment effect.

Spillover is a possibility that can thus bias straightforward estimates of many conventional treatment effect parameters.[51] One could turn around and argue that the spillover informed treatment effect

$$E\left(Y^1\right)-E\left(Y^0|S\right)$$

is somehow of interest. Remembering that the ultimate goal of program impact estimation is to understand the influence of programs on human welfare so that we can focus efficiently scarce resources, we would then be confronted with an important question: if spillover effects are meaningful enough that we would want them to inform an impact parameter such as this, then don't we want somehow to estimate them to obtain the fullest possible picture of program impact?

The problem is that, at least within the confines of our cluster, there is no way to do that. The spillover effect is

$$E\left(Y^0|S=1\right)-E\left(Y^0|S=0\right)$$

Unfortunately, we simply do not observe a subpopulation in that cluster whose experiences allow us to estimate $E(Y^0|S=0)$.

One possible remedy for this is to perform "double randomization" (Angelucci & Di Maro, 2015). Although this sounds really fancy, the basic idea is pretty simple. With a cluster-based data

[51] For instance, by similar reasoning it is not hard to see that spillover effects can preclude simple unbiased estimation of the average effect of treatment on the treated.

design in mind, we first randomize some clusters to receive the program and others not to do so. We then randomize participation status P in those clusters randomized to receive the program at all. As we have seen from the conversation to this point, randomization within clusters that receive the program at all solves the problem of selectivity but does not address bias to estimates of common impact parameters from spillover. However, randomization of some clusters not to receive the program at all creates a subpopulation for which we actually do observe

$$E\left(Y^0\right)$$

We could then estimate the average treatment effect

$$E\left(Y^1 - Y^0\right) = E\left(Y^1\right) - E\left(Y^0\right)$$

by estimating $E(Y^1)$ with a (representative) sample average of the observed outcome for participants (which necessarily can be formed only with those clusters randomized to receive the program at all) and by estimating $E(Y^0)$ with a (representative) sample of individuals from clusters randomized not to receive the program.

However, that is not all we could estimate. We could also estimate spillover. A valid estimate of spillover

$$E\left(Y^0|S = 1\right) - E\left(Y^0|S = 0\right)$$

can be formed by comparing observed outcomes for non-participants in clusters randomly assigned to receive the program and observed outcomes for everyone in clusters randomly assigned not to receive the program. A test for spillover effects could be as simple as a test of the difference in these two means.

A trivial extension of this design would allow us to understand how spillover depends on participation levels. All that would be required is, in the clusters randomly assigned to receive the program at all, to randomly vary the proportion of the population randomly assigned to participate. We thus might consider circumstances were 5, 10, 20, 25, 50 and 75 percent of the population are assigned to participate. We could then estimate spillover for each type of cluster receiving the program to get some sense of how spillover depends on the proportion of the population participating in the program.

One could then even form some rough estimate of the overall impact of the program at the population level. Specifically, impact for the typical individual in program areas would be

$$P_P \cdot E\left(Y^1 - Y^0\right) + (1 - P_P) \cdot \left\{E\left(Y^0|S = 1\right) - E\left(Y^0|S = 0\right)\right\}$$

where P_P is the percentage of the population participating in the program. This would need to be calculated separately for each different type of program cluster in terms of the percentage of the cluster population assigned to participate. This would be a natural approach, for instance, for considering overall effectiveness for a cost-effectiveness analysis of the program.

For the purpose of projecting total program impact “in the wild” (that is, when the program is operating under normal circumstances after the double randomization experiment has concluded) this is a bit of a shaky estimator since in normal operational circumstances (such as outside of the confines of a randomized controlled trial) it seems likely that individuals would select themselves into participation status, implying some rearrangement of the types of individuals experiencing the program directly or indirectly via spillover effects. To some extent (to the extent that selection into participant status is driven by observables) this could potentially be addressed via a more elaborate model of P_P than the simple proportion participating.

This entire scheme depends on the idea that the residents of those clusters randomly assigned not to receive the program experience no spillover from those that were randomly assigned to receive it. This is probably in some contexts a questionable assumption. One would do well to remember the interrelatedness of all things before dismissing this possibility. Indeed, the distant and tenuous but critical connection is the foundation for the entire idea of the “butterfly effect.”[52] Looked at from the philosophical standpoint of the butterfly effect, there is probably no one on earth that we can be sure has not experienced some spillover from a given program. This might seem a ridiculous proposition until one considers the theory that everyone on earth is separated by at most six interpersonal connections. The double randomization estimator thus in practice really relies on the idea that whatever spillover does occur is negligible (whatever “negligible” means in practice).

Double randomization is sometimes a possibility. When it is, the challenge of the sampler is pretty simple: to collect representative samples of participants and non-participants from clusters randomly assigned to receive the program and a representative sample from the overall population of clusters randomly assigned not to receive the program. The challenge is then to insure that the samples are large enough to detect the program impact and spillover effect that the study goals require to be detected.

Unfortunately, in our experience double randomization is often not a possibility. Indeed, rarely are we presented with the opportunity to randomize program presence at either level required under the double randomization scheme. Typically, in fact, we are approached to conduct an evaluation only *after* many of the parameters of program allocation across communities/clusters have already been established. Specifically, communities/clusters to receive the program have often already been determined. Moreover, for any number of reasons within-program cluster randomization of participation is often not desirable to program managers, the funders of the evaluation, or both. Alternatively, randomization might not be feasible.

Nonetheless, some of the basic conceptual logic of the double randomization case often informs evaluation design discussions even when randomization is not possible. The potential for bias due to spillover is still present. The only difference is that now it is augmented by the possibility of

[52] This “butterfly effect” is basically the idea that small causes can have large effects. This has roots in mathematics extending back at least as far as Poincaré. It was perhaps most colorfully illustrated by Ray Bradbury’s 1952 short story “A Sound of Thunder,” in which the death of a butterfly in the late Jurassic influences events in the mid-21st century.

selectivity bias since randomization of participation within communities that do receive the program is no longer possible. When such within-community randomization of participation is not possible participants sort themselves into participant and non-participant status, with some constraint on that sorting process presented by enrollment eligibility criteria that the program might impose.

It is our experience that spillover and selectivity are confronted somewhat independently in such cases. A typical response to spillover is to try to seek control/comparison subsamples for the purpose of estimating some parameter associated with the outcome in the absence of participation Y^0 (such as $E(Y^0)$, $E(Y^0|P = 1)$, or the implicitly no-participation counterfactual in some more elaborate causal model) from clusters that do not receive the program.

The basic motivation for this is much the same as in the double randomization case: it is assumed that those not residing in communities/clusters receiving the program are not likely to experience spillover from program operations or the experiences, circumstances, or behaviors of participants. There is, however, a crucial problem introduced by the failure to randomize clusters for eligibility to receive the program. Randomization of clusters/communities in the double randomization case allowed us to be certain that the population of clusters assigned to receive the program were the same on average as those assigned not to do so. Let

$$A = \begin{cases} 1 & \text{if a cluster is assigned to receive the program} \\ 0 & \text{otherwise} \end{cases}$$

The first stage of randomization thus insured that

$$E\left(Y^1, Y^0|A = 1\right) = E\left(Y^1, Y^0|A = 0\right) = E\left(Y^1, Y^0\right)$$

where the term to the right of the second equality is the expectation over the populations of all of the clusters (i.e., those eligible and not eligible to receive the program) involved in the double randomization. It was this that allowed us to estimate impact parameters associated with Y^0 (such as $E(Y^0)$) using the observed outcomes of those in clusters randomly assigned not to be eligible for/not to receive the program.

The trouble is that when the first stage of the double randomization process does not happen we cannot guarantee that

$$E\left(Y^1, Y^0|A = 1\right) = E\left(Y^1, Y^0|A = 0\right) = E\left(Y^1, Y^0\right)$$

will hold. Essentially, the process of determining which community/cluster recieves the program is no longer random. Instead, it probably reflects endogenous program placement (Rosenzweig & Wolpin, 1986; Pitt, Rosenzweig, & Gibbons, 1993; Gertler & Molyneaux, 1994; Angeles, Guilkey, & Mroz, 1998, etc.). For instance, programs might locate to the most under-served areas where the perceived need for them is greatest. This implies possibilities such as

$$E\left(Y^0|A = 1\right) < E\left(Y^0|A = 0\right)$$

On the other hand, the cynical expectation is that, sometimes, programs might decide to operate where conditions are favorable (and hence where it is easier to operate), implying that

$$E\left(Y^0|A=1\right) > E\left(Y^0|A=0\right)$$

This is a kind of selectivity bias between clusters.

Now let us bring the conversation full circle to spillover. Suppose that we wish to estimate the average treatment effect

$$E\left(Y^1 - Y^0\right) = E\left(Y^1\right) - E\left(Y^0\right)$$

where $E(\cdot)$ is with respect to the overall population across the communities/clusters that are eligible ($A = 1$) and ineligible ($A = 0$) for the program. Clearly, the absence of randomization of participation within communities/clusters receiving the program means that estimation of $E(Y^1)$ is no longer straightforward. Due to self-selection (perhaps constrained by eligibility criteria that may or may not be constant or transparent so that the evaluators can observe and model them) we can no longer be sure that

$$E\left(Y^1|P=1, A=1\right) = E\left(Y^0\right)$$

due to selection at the individual level within eligible clusters as well as selection across clusters for eligibility to receive the program. Nothing short of double randomization can remove this "fog of selectivity."

However, the focus of the spillover concern is usually with estimating the counterfactual parameters such as $E(Y^0)$. We can think of several layers of bias. First, were we to confine our evaluation to the clusters eligible for the program, the bias to our estimate of the counterfactual $E(Y^0)$ associated with, for instance, a simple average of observed outcomes Y from representative samples of non-participants would be

$$E\left(Y^0|P=0, S=1, A=1\right) - E\left(Y^0\right)$$

Second, were we to select ineligible clusters (presumably) not subject to spillover concerns and use a simple average of the observed outcomes Y for representative samples of the populations within them, the bias would be

$$E\left(Y^0|P=0, S=0, A=0\right) - E\left(Y^0\right) = E\left(Y^0|A=0\right) - E\left(Y^0\right)$$

where the second equality emerges simply because ineligibility of one's cluster for the program precludes the possibility of spillover or participation as an individual (or household, or any other unit below the cluster level).

There are two "moving parts" between these two bias measures. In the first case, we would experience bias due to spillover and within-cluster selectivity, but not from the non-participant sample being drawn from different cluster-level background circumstances than participants. In the second case we would still experience bias due to within cluster selectivity but not due to spillover. We would, however, experience bias from the comparison populations being drawn from (possibly very) different cluster contexts.

Which bias is worse? That is simply impossible to say a priori with cross-sectional data. Perhaps something could be learned with longitudinal data with a pre-program panel from clusters where $A = 1$ and $A = 0$.

The big takeaway from this discussion is to be careful about reflexively addressing spillover through a design where non-participant comparison samples are drawn from clusters in which the program does not operate. By following this route one *may* reduce spillover bias. However, we likely introduce another bias: the comparison population is drawn from (possible very) different community environmental circumstances, potentially calling into question the degree to which their observed outcome ($Y = P \cdot Y^1 + (1-P) \cdot Y^0 = 0 \cdot Y^1 + (1-0) \cdot Y^0 = Y^0$) serves as a useful indication of what might have happened in program areas in the absence of the program.

There is no clear-cut guidance that we can offer for sampling in the case where there might be spillover but double randomization is not in order. We have seen it argued (and strongly) that estimates of the counterfactual no-participation outcome should be sought from populations not proximate to the area of program operations. The motivation for this is to avoid contamination from spillover. However, as this discussion has revealed, in pursuing this approach we might simply be swapping one bias for another, with little prior understanding of which bias is actually larger.

One way of potentially allaying concerns regarding this can be found in a pre-matching strategy. This would involve attempting to use observed characteristics to find a set of potential control clusters that most closely (in whatever sense is deemed important given the assumed behavioral pathways for the outcome of interest Y for the evaluation) resemble the program clusters. This matching could conceivably be done either before or after selection of comparison clusters. Conceptually, it is probably more persuasive to match before selecting control clusters (so that random selection of clusters occur across the non-program clusters that match the program clusters). This might not be feasible in all cases, however, if the universe of potential control clusters is so large that it would be logistically difficult to gather matching information from all of them.

Chapter 3

Sample Size *Guesstimation*

In this chapter, we discuss sample size estimation, or rather sample size *guess*timation. We are not trying to be funny with the term *guess*timation. Sample size estimation for impact evaluation, as well as virtually every other kind of goal for a population survey, is essentially the scientific formalization of an educated guess about required sample size. Typically, (often strong) assumptions need to be made about likely features of the data to be collected. In this context, "assumptions" is simply another way of saying "guesses."

Sample size estimation itself typically begins with formal analytical formulas for computing sample size estimates. These formulas are varied, with different formulas appropriate to different survey objectives. For instance, different impact evaluation estimation methods require different sample size estimation strategies.

Despite their diversity, the formulas tend to have four things in common. First, they yield estimates of required sample size in terms of the unit of analysis for the eventual impact evaluation (as opposed to being in terms of the unit of sample selection upon which much budgeting and planning and/or primary sampling unit selection must be done). For instance, we might wish to assess impact among children under age five from poor households. The sample size estimate provided by the formula will be the required sample size in terms of these units of estimation. However, it is very unlikely that we will as investigators possess a list of poor children under five years old for sample selection purposes. Instead, we will typically engage in multistage selection, beginning with some administrative unit such as the cluster, ward, etc. serving as the primary sampling unit.

Translating the sample size of poor children provided by the sample size formula into a sample size of primary sampling units will require assumptions. First, in a multistage sample selection process, we must know how many sampling units we can expect to select per each sampling unit selected at the next highest level of selection. For instance, how many poor children under five can we expect to find per household selected?

Second, the sample size estimation formulas tend to assume simple, independent selection with equal probability of selection and no stratification (beyond the participant/non-participant dichotomy) or clustering. In particular, they tend to involve an assumed sampling variance based on this simple sample selection process. We must somehow correct the simple sample size estimate to take into account the selection design (clustering, unequal selection probabilities, etc.) we will actually use, which will leave us with a different sampling variance than the one based on simple, independent, equal probability selection on which the sample size formula was based. This typically requires some assumptions about the variance cost of our design, typically expressed through a design effect.

Third, in many instances these sample size formulas tend to require assumptions about the likely true population value of the impact parameters, or components of the impact parameters, that we wish to estimate. In other words, appropriate sample size for estimating impact can depend on the same particulars behind true population impact, which is the very thing we wish to learn about through survey and estimation. Indeed, an old joke about sample size estimation is that in many applications, if one is completely sure of the sample size for the survey then they don't really need to conduct the survey!

Fourth, the actual estimation formulas themselves often rely on approximations. For instance, some formulas set the stage for solution for the sample size by defining either the probability that the eventual estimate of the population parameter will be within an interval of sampling variation of some size or the probability that the sample size will be large enough to detect a difference in parameter values between two populations (the so called "power function"). Notice that in both cases the departure point is a probability statement. However, to go from that to a solution for a sample size typically requires some kind of distributional assumption. Normality is a common one. Whether it is always a good one is another matter.

The bottom line is that assumptions/guesses feed every aspect of the sample size estimation process. This means that sample size estimates can be quite misleading when the assumptions/guesses are wrong. This would be an argument against conventional sample size estimation practice if not for the absence of any alternative approach to sample size determination not virtually guaranteed to be even more unreliable.

Given the potential unreliability of sample size estimation, those designing impact evaluations need to understand the consequences of under- or over-estimation of required sample size. Against the departing goals of the impact evaluation (in general, to be able to detect a certain degree of impact with a certain degree of probability) under-estimation of sample size means that these departing goals cannot be met by the sample collected. Over-estimation of the sample size implies that more resources were dedicated to the survey than necessary.

To fix ideas, consider a sample size estimate of 1,000 units (this could be thought of either in terms of units of estimation or primary or other sampling units). If it turns out that the required sample size was 800 units (in other words, upon collection of the sample it becomes apparent that this was likely a more appropriate sample size given the goals of the impact evaluation), then resources required at the margin to collect 200 units would have been wasted (from the strict standpoint of achieving the goals of the evaluation). On the other hand, if the required sample size is 1,200, or 200 units more than collected, then *all* of the resources dedicated to the evaluation have been wasted in the sense that there has been a failure to achieve the motivating estimation goals of the evaluation.

Looked at from this standpoint, the cost of over-estimation of sample size is probably less than the cost of under-estimation. Given that sample size estimation is awash in assumptions/guesses, this suggests that one should perhaps make reasonably conservative assumptions, where by "conservative" we mean assumptions that would tend to lead to a larger final sample size estimate. Admittedly, resources are scarce and one does not wish to waste them. Nonetheless, it must always be remembered that when we cut corners in sample size estimation by favoring assumptions that would lead to smaller sample sizes simply because they would lead to a smaller estimated sample size, we might actually be wasting resources in the end.

The reliance on so many assumptions/guesses is also somewhat liberating. The fact that sample size estimation is in some sense a castle built on sand should give one pause about worrying too much about some arcane consideration about a specific aspect of the sample size estimation process: Given all of the strong assumptions that need to be made for sample size estimation (i.e., given the uncertainty that surrounds virtually every aspect of the process), less probably rides on one particular arcane consideration than one might, caught up in the intensity and focus of the moment, think. Moreover, in many cases where one faces alternative options at such decision points it is typically not obvious which one will yield the more appropriate sample size estimate.

This statement is not a call for deliberate sloppiness in sample size estimation, but instead a suggestion that sample estimation is a good example of the old advice not to make "the perfect the enemy of the good." Rather than spending incredible amounts of time focused on one specific, detailed consideration (and in the process likely slowing and complicating the sample size estimation process and driving up design costs[1]) one might instead simply adopt what appears to be a (reasonably) more conservative course for that aspect of the sample size estimation process.

3.1 The Classic Sample Size Estimator

We now introduce and discuss a classic and basic approach to sample size estimation that has been applied to determine sample sizes for a wide variety of population surveys with various objectives. This will provide some feel for the conceptual logic motivating sample size estimators.

That said, the importance of the classic models discussed in this section really cannot be overstated. The formulas that we discuss for detection of differences in population parameters across subpopulations (such as those who participate in a program and those who do not do so) have, in the authors' experience, been a workhorse one for sample size estimation for impact evaluations to which they have contributed as sampling consultants.

There are a number of reasons for this. First, at their core most evaluations involve, essentially, the comparison of some outcome of interest between program participants and non-participants. This estimation approach is central to randomized controlled trials (RCTs).[2] Even when random-

[1] If nothing else through staff time for sampling!

[2] Even when quasi-random estimation approaches, such as multiple regression, are applied to experimental samples (i.e., those in which program participation is determined purely by random assignment).

ization of program participation is not a possibility, a notional RCT is often the idealized lens through which the design of the impact evaluation is framed. For instance, the framework of an idealized experimental sample[3] is a useful intellectual departure point for thinking about the identification challenges and possibilities presented by the circumstances surrounding the design and circumstances of an evaluation. Most quasi-experimental estimators are also implicitly or explicitly motivated by the idea of finding comparable comparison groups for the purpose of estimating differences in outcomes between our participant and comparison samples so operationalized.

Moreover, the eventual quasi-experimental estimators to be used for estimation of causal program impact (either in the face of an *ex-ante* plan to rely on quasi-experimental methods or in the event of unsuccessful randomization in an ostensible RCT) might not be apparent at the design phase of an evaluation. Quasi-experimental estimators often rely on tests of their appropriateness (such as overidentification tests) the outcome of which cannot be known at the design phase of an evaluation. It is therefore sometimes difficult to know which among several potential quasi-experimental strategies will ultimately be pursued. Even when there is only one feasible or obvious choice, projecting the actual performance or credibility of a quasi-experimental estimator from the design phase can involve many strong assumptions (including instrument strength, degree of observed "within" variation, or the extent of failure of common support).[4] In the face of this uncertainty, the relatively simple, classic sample size estimators, motivated by explicit comparison of observed outcomes between participant and comparison or control samples, provide a sort of "reality check" about whether proposed sample sizes would be sufficient even for the comparatively straightforward estimation under a basic RCT. If the sample size is insufficient for the relatively simple comparison of means, then it is often probably optimistic to expect that it will be sufficient for a more elaborate quasi-experimental estimation approach.

We thus begin with a basic, classical approach to sample size estimation, one of great importance to impact evaluation but also applied to surveys with many other sorts of goals. To help the reader through this process, we ease into this gradually.

Toward that end, we begin with a sample size estimation goal far simpler than the comparison of a parameter across different populations (such as between participant and non-participant subpopulations) that most directly reflects the basic challenge of impact evaluation. Instead, we begin by considering a much simpler goal: estimating a parameter value for a population with some

[3] Lance, Guilkey, et al. (2014) define an experimental sample as one in which program assignment (to participant and non-participant subsamples) is fully random. In other words, an experimental sample is one yielded by a *successful* RCT. It can also sometimes emerge due to a natural experiment that generates full randomization of participation (though when one thinks of natural experiments the focus is usually on less than full randomization, as in the case of an instrument that delivers a random channel of variation in participation even as participation is on the whole non-randomly determined). Non-experimental data would then be data within which program participation is not random, either due to the failure of an ostensible RCT design, an incomplete natural experiment (in the sense that it did not fully randomize participation) or because randomization of participation was not feasible as a design option from the outset. In the text we tend to think of the latter possibility: that randomization was not an explicit feature of the design of the impact evaluation, and its identification strategy anticipated the need for quasi-experimental identification of program impact.

[4] *A priori* uncertainty about the particulars of quasi-experimental designs is an argument for study designs that are robust in the sense that they could potentially meet the data needs for numerous quasi-experimental methods.

desired degree of precision.

3.1.1 Precision

Let $\overline{Y}$ be some population parameter (the value of which we might wish to estimate through a survey). We use the overline "$\overline{\ }$" to suggest that this parameter is an average. In other words, as in the last chapter, the motivation for our discussion is the idea that $\overline{Y}$ is some sort of population average. This is a natural entry point for thinking about sample size estimation for program impact evaluation since we are usually trying to estimate some kind of average (such as an average treatment effect, average effect of treatment on the treated, etc.).

To fix ideas, suppose for now that $\overline{Y}$ is a proportion. For instance, $\overline{Y}$ could be the modern contraceptive prevalence rate, the proportion of children under five who are stunted, the percent of teenagers who report being sexually active, etc. We will denote our estimate of $\overline{Y}$ by $\hat{\overline{Y}}$. Because $\hat{\overline{Y}}$ is an estimate, it will almost certainly differ from sample to sample from the true parameter value $\overline{Y}$ to some degree (that is, there will be "sampling error"), even if sample selection and estimation lead to unbiased estimates.

Suppose that we wished to control the degree of sampling error. For instance, suppose we wanted sampling errors to be such that the difference between our estimate $\hat{\overline{Y}}$ and the population parameter $\overline{Y}$ would be no more than d in absolute value:

$$\left|\hat{\overline{Y}} - \overline{Y}\right| \leq d$$

where $\left|\hat{\overline{Y}} - \overline{Y}\right|$ indicates the absolute value of $\hat{\overline{Y}} - \overline{Y}$. We are thus suggesting that we want to limit sampling errors to d in size (d is also sometimes referred to as a sampling error margin, error margin, margin of error, etc.).[5]

Of course, we can never *guarantee* that the sampling error will be less than d as long as $\overline{Y} + d$ is less than 1 or $\overline{Y} - d$ is greater than zero: since we are talking about a proportion, there is likely some positive probability (however small) that our estimate $\hat{\overline{Y}}$ could take on any value between 0 and 1. What we can do, however, is to set a sample size that insures the *probability* of an error larger than d is no more than some tolerable level. Thus, what we could do is insure that the probability that the error $\hat{\overline{Y}} - \overline{Y}$ is less than or equal to d is no less than $1 - \alpha$.

In other words, we can set a sample size such that

$$Pr\left(\left|\hat{\overline{Y}} - \overline{Y}\right| \leq d\right) = Pr\left(-d \leq \hat{\overline{Y}} - \overline{Y} \leq d\right) \geq 1 - \alpha$$

where $Pr(\cdot)$ denotes the probability of an expression.

In sampling discussions, α is the *significance level*. What we are saying is that we want the probability that the absolute value of the sampling error is less than or equal to d to be at least $1 - \alpha$.

[5] And, in the language of the last chapter, what we are really talking about here is limiting the extent of sampling variation.

Recall from basic statistics that the confidence interval for the estimate $\hat{\overline{Y}}$ is

$$\hat{\overline{Y}} \pm z_{1-\frac{\alpha}{2}} \cdot se\left(\hat{\overline{Y}}\right)$$

where $se(\hat{\overline{Y}})$ refers to the standard error of $\hat{\overline{Y}}$ and $z_{1-\frac{\alpha}{2}}$ is the critical value from the standard normal distribution such that the probability that a standard normal random variable is greater than that value is $\alpha/2$. Similarly, $-z_{1-\frac{\alpha}{2}}$ is the critical value from the standard normal distribution such that the probability that a standard normal random variable is less than that value is $\alpha/2$. This is shown in Figure 3.1, which illustrates the standard normal probability density. The shaded areas represent the probabilities that a standard normal random variable is less than -2 or greater than 2: Each of the shaded areas is $\alpha/2$ in size, and the two together are α in size.[6]

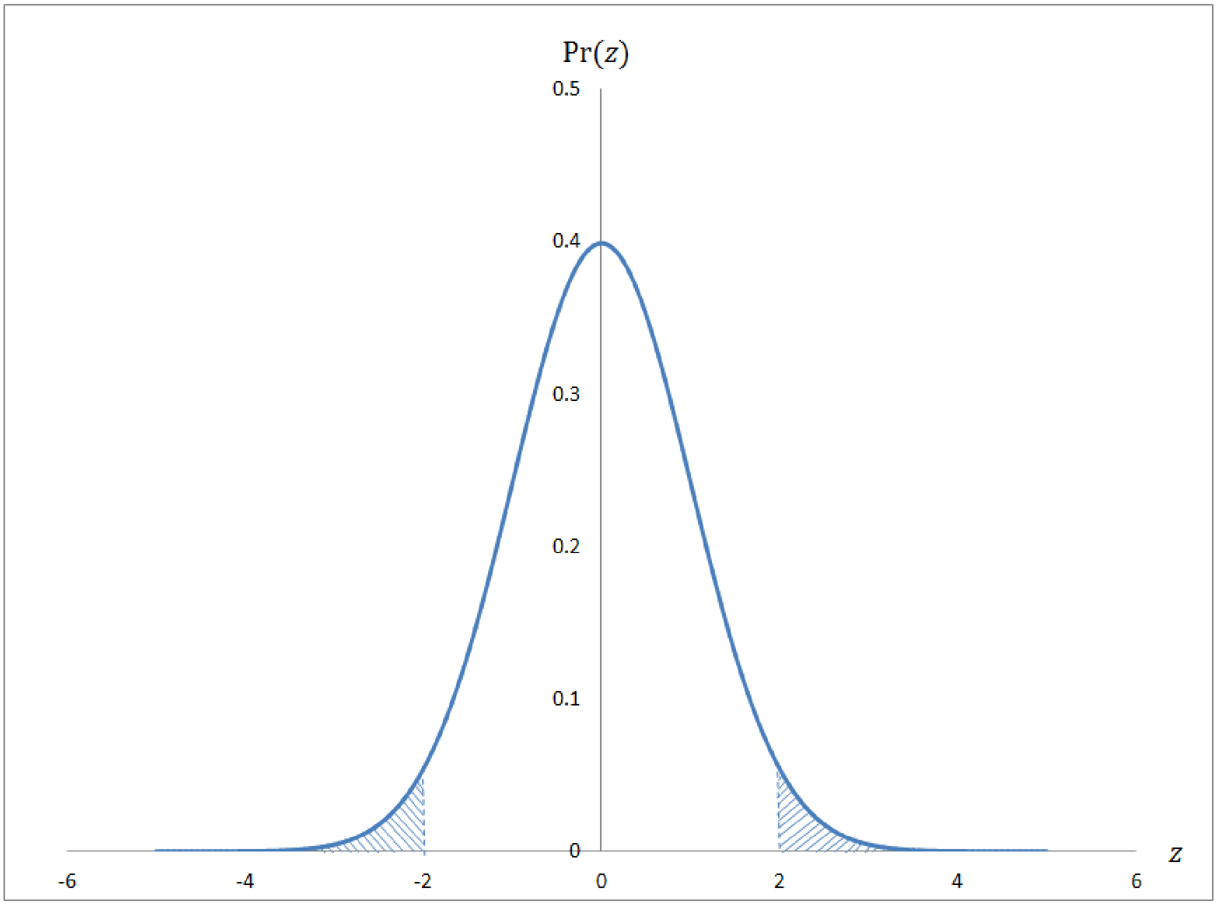


Figure 3.1. Critical values of the normal density

It is tempting to say that there is a $1-\alpha$ percent chance of $\overline{Y}$ being in the confidence interval but this is technically incorrect: $\overline{Y}$ is not a random variable in this context but instead a parameter and as such not subject to variation that can be characterized through probabilities. For instance,

[6] The focus on the critical values -2 and 2 in Figure 3.1 was mostly for visual simplicity, but not entirely. A commonly pursued value for α is 0.05, for which the critical values from the standard normal distribution are approximately -1.959964 and 1.959964.

the contraceptive prevalence rate for the population of women ages 15–49 is not a random variable. It is a parameter with, in principle, some actual, true specific value across that population.[7] Instead, because it relies on estimates $\hat{\overline{Y}}$ and $se(\hat{\overline{Y}})$ it is the interval boundaries themselves that are random. Hence, the proper interpretation is to say that $100 \cdot (1 - \alpha)$ percent of all of the intervals constructed in that way will contain the true parameter value $\overline{Y}$. For present purposes, the width of the confidence interval can be thought of as a gauge of how uncertain we are about the unknown parameter $\overline{Y}$.

The key step in developing a sample size estimate is then to use the confidence interval, being a gauge of uncertainty, as a way to characterize our tolerable sampling error. Specifically, we insure that the tolerable bounds for the sampling variation in the estimate is the size of the confidence interval by the following:

$$\hat{\overline{Y}} \pm d = \overline{Y} \pm z_{1-\frac{\alpha}{2}} \cdot se\left(\hat{\overline{Y}}\right)$$

Subtracting $\overline{Y}$ from both sides we have

$$d = z_{1-\frac{\alpha}{2}} \cdot se\left(\hat{\overline{Y}}\right)$$

Squaring both sides, we have

$$d^2 = z^2_{1-\frac{\alpha}{2}} \cdot var\left(\hat{\overline{Y}}\right)$$

where $var(\hat{\overline{Y}})$ is the variance of $\hat{\overline{Y}}$ (recall that the standard error is the square root of the variance of $\hat{\overline{Y}}$, and hence the square of the standard error is the variance of $\hat{\overline{Y}}$).

The next step involves recognizing that the variance of a random variable (such as the estimate $\hat{\overline{Y}}$) essentially always involves somehow the number of observations n. In the case of a proportion (which ultimately depends on a binary outcome), the variance is

$$\frac{\overline{Y} \cdot \left(1 - \overline{Y}\right)}{n}$$

Unfortunately, we cannot observe $\overline{Y}$, and so use our prior guess about $\hat{\overline{Y}}$, which is tantamount to a prior guess about $\overline{Y}$ (more on this below). This implies that the estimator of the variance of $\overline{Y}$ is

$$var\left(\overline{Y}\right) = \frac{\hat{\overline{Y}} \cdot \left(1 - \hat{\overline{Y}}\right)}{n}$$

We then have

$$d^2 = z^2_{1-\frac{\alpha}{2}} \cdot \frac{\hat{\overline{Y}} \cdot \left(1 - \hat{\overline{Y}}\right)}{n}$$

Or, solving for n, we have

$$n = \frac{z^2_{1-\frac{\alpha}{2}} \cdot \hat{\overline{Y}} \cdot \left(1 - \hat{\overline{Y}}\right)}{d^2}$$

[7] Though as we have seen, the "super-population" approach to this can muddy the waters a bit on this point.

where n is the required sample size for the survey. This is a simple sample size estimation formula in part because it was motivated by a simple sampling goal (to try to control sampling error). But note that even in this simple case the foundation of the sample size estimate was a well-stated survey goal (to learn about $\overline{Y}$ with some tolerable level of sampling error d).

Although this was a simple example, it has elements common to many sample size estimation methods. These include:

1. It requires a well-stated survey goal. This usually involves clarifying the precision or degree of certainty with which we wish to learn something about a population (or different populations), and then specifying the probability (or probabilities) that we will indeed stay within those bounds of precision or degree of certainty.

2. Making some prior assumptions about parameters, such as a plausible value for the estimate $\hat{\overline{Y}}$ (more on this in the paragraph to follow), is necessary.

3. We needed to make some kind of distributional assumption (in this case involving the standard normal distribution) to allow us to characterize the probabilities in some concrete way in order to derive a sample size estimate.

4. The variance of the estimate was more or less a central building block to deriving a sample size estimate. This is why it is sometimes suggested that "sample size estimators are variance based." Certainly, that is more or less the case with the classical variance estimators that we consider in this section.

While the world of sample size estimation can (quickly) get more complicated than this, surprisingly many sample size estimators implicitly or explicitly rest on these basic conceptual pillars.

Implementing this sample size estimator (i.e., actually calculating a sample size with it) in the case where $\overline{Y}$ is a proportion requires making some assumption about the likely value of the estimate $\hat{\overline{Y}}$. The assumed value will influence the required sample size through the term in the numerator $\hat{\overline{Y}} \cdot (1 - \hat{\overline{Y}})$. At the root of this is the fact that $\hat{\overline{Y}} \cdot (1 - \hat{\overline{Y}})$ is the numerator of the variance, and the sample size depends positively on the variance.

Put simply, the larger $\hat{\overline{Y}} \cdot (1 - \hat{\overline{Y}})$ is, the larger will be the required sample size. So the sample size in some sense rests on our prior guess about the estimate of the population proportion, $\hat{\overline{Y}}$.

Figure 3.2 illustrates the value of $\overline{Y} \cdot (1 - \overline{Y})$ for a proportion $\overline{Y}$ in general. As the figure illustrates, the value of $\overline{Y} \cdot (1 - \overline{Y})$ depends heavily on $\overline{Y}$, with a maximum value achieved at $\overline{Y} = 0.5$. This is, among other things, an excellent example of how estimated sample size can depend on prior assumptions about key inputs to the sample size formula (in this case the likely value of the estimate of the population proportion $\hat{\overline{Y}}$).

What constitutes a good *ex-ante* guess for $\hat{\overline{Y}}$? Well, although $\hat{\overline{Y}}$ is an estimate, it is hoped that it is an unbiased one. Given this, in most applications an *ex-ante* or prior guess about $\hat{\overline{Y}}$ is tantamount to a prior guess about the actual population parameter $\overline{Y}$. A good guess would then be a (reasonably informed) guess about the true value of the population parameter as of the time of the survey.

From the formula

$$n = \frac{z^2_{1-\frac{\alpha}{2}} \cdot \hat{\overline{Y}} \cdot \left(1 - \hat{\overline{Y}}\right)}{d^2}$$

we can see some other likely general patterns to sample size estimates. To begin with, the sample size will clearly increase as the desired sampling error d shrinks. This makes sense: d is the most direct indicator of our desired degree of precision in estimation, and more precise estimation in terms of a smaller d requires more information in the form of observations. The sample size will also increase as α gets smaller. This also makes intuitive sense: More observations will be required to increase the probability that estimates will stay within the sampling error.

That said, one has to be careful about predicting the impact on required sample size from changes in assumed $\overline{Y}$, d or α. For one thing,

$$n = \frac{z^2_{1-\frac{\alpha}{2}} \cdot \hat{\overline{Y}} \cdot \left(1 - \hat{\overline{Y}}\right)}{d^2}$$

is a nonlinear formula. Therefore, predicting the sample size implications of a change in one of these parameters, even holding the other two constant, can be tricky, as the magnitude of the impact of a

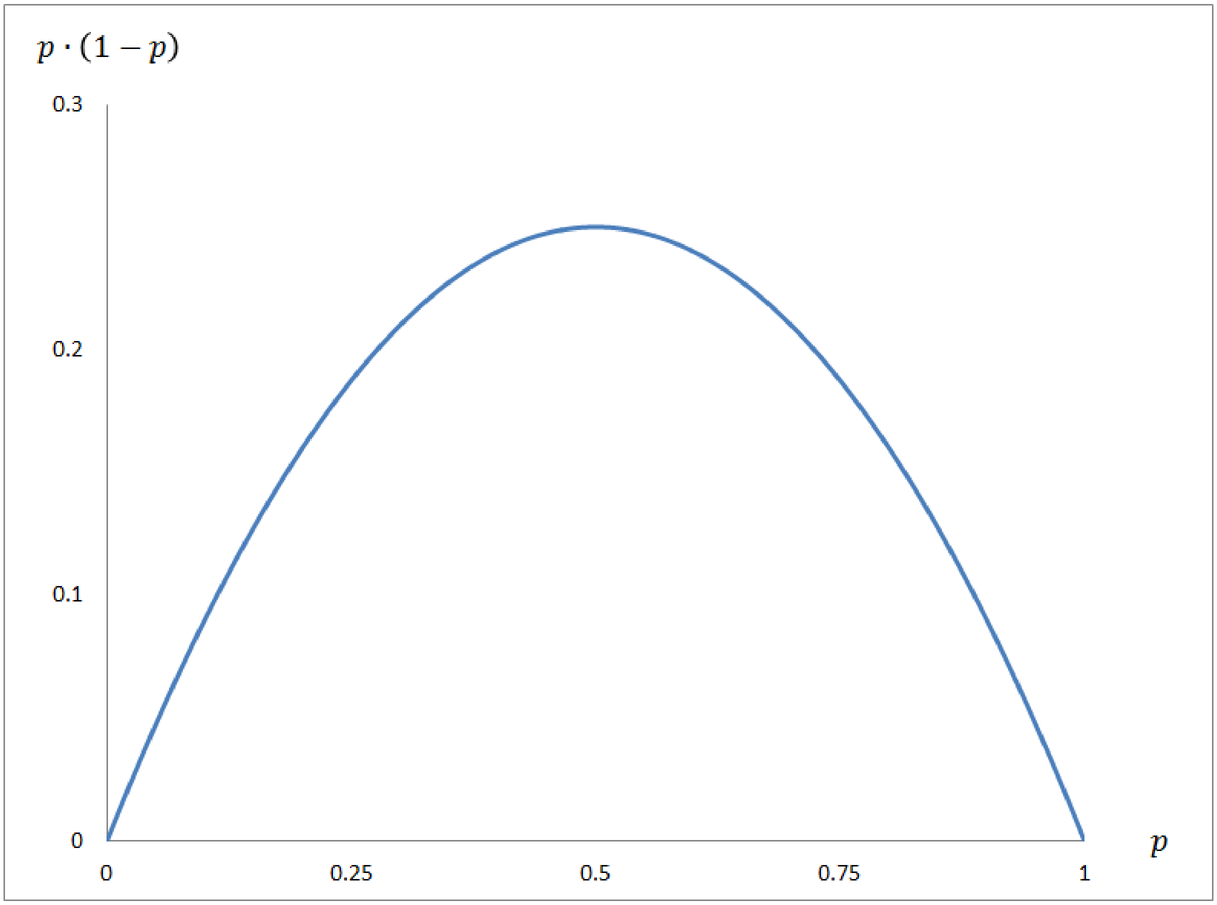


Figure 3.2. How the proportion value $\overline{Y}$ influences $\overline{Y} \cdot (1 - \overline{Y})$

change in one of them can depend on the values of the other two, and in complex ways. Moreover, predicting the net impact of simultaneous changes in more than one of these parameters is very dangerous: one should always actually perform the calculations to see what the net impact of the changes will be.[8]

To perhaps make this complexity a bit more concrete, let us consider some specific numerical examples. Suppose, for instance, that we wish to estimate the contraceptive prevalence rate for women ages 15–49 through a survey. We can look to past estimates, and a reasonable (or at least not wholly implausible!) assumption about the trend in that rate is to guess that, as of the time of our survey, it might be 0.3. We would thus assume that the most likely value for our estimate $\hat{\overline{Y}}$ would be 0.3. Suppose as well that we wish to have an error $\left|\overline{Y} - \hat{\overline{Y}}\right|$ no greater than 15 percent of the true population parameter value. Given our best prior of $\overline{Y}$ (and thus $\hat{\overline{Y}}$) of 0.3, this implies an error d of

$$0.15 \cdot 0.3 = 0.045$$

Finally, we decide that we want the probability of an error larger than this being no larger than 5 percent, implying an α value of 0.05 and a critical value of

$$z_{1-\frac{\alpha}{2}} = z_{0.975} \approx 1.959964$$

The sample size estimate is then

$$n = \frac{(1.959964)^2 \cdot 0.3 \cdot (1 - 0.3)}{(0.045)^2}$$

$$= 398.37351 \approx 399$$

We would need a sample of around 399 women to estimate the contraceptive prevalence rate.[9]

Suppose that we wished to be more precise in our estimate. If we wanted to have an error that was only 5 percent of the population parameter value ($0.05 \cdot 0.3 = 0.015$), the new sample size is

$$n = \frac{(1.959964)^2 \cdot 0.3 \cdot (1 - 0.3)}{(0.015)^2} = 3585.36157 \approx 3,586$$

In other words, reducing the value of the error by 67 percent leads to a nearly nine fold increase in the sample size. In direction, the change in required sample size intuitively reflects "statistical fundamentals": if we want more precision to our estimate we will need more information, in the form of additional observations, to form that estimate. That said, could the reader have guessed intuitively correctly the *magnitude* of the change in the required sample size? The authors almost certainly could not have done so.

Returning to the original example, let us consider some other variations. To begin with, let us

[8] This is one reason why the authors essentially always resist requests to make "off the cuff" guesses (for instance, in the course of meetings) about the likely sample size implications of proposed changes in sampling parameters.

[9] A common practice in sample size estimation is to round up a fractional sample size estimate to the nearest integer.

consider a few different prior assumptions regarding $\overline{Y}$ (that is, $\hat{\overline{Y}}$). Suppose that our prior guess for $\overline{Y}$ (and $\hat{\overline{Y}}$) is 0.35 instead of 0.3. d is then

$$0.15 \cdot 0.35 = 0.0525$$

The sample size would then be

$$n = \frac{(1.959964)^2 \cdot 0.35 \cdot (1 - 0.35)}{(0.0525)^2}$$

$$= 317.0728$$

Suppose instead that the prior guess was 0.25. d is then

$$0.15 \cdot 0.25 = 0.0375$$

The sample size would then be

$$n = \frac{(1.959964)^2 \cdot 0.25 \cdot (1 - 0.25)}{(0.0375)^2}$$

$$= 512.19452$$

One could reasonably ask how these results could be consistent with the notion (from Figure 3.2) that the required sample size should grow the closer is the prior assumption about $\overline{Y}$ to 0.5. The answer is that the prior assumption is not the only moving part in these examples: we have also changed the the sampling error d. In these examples, this has overwhelmed the normal consequences of moving the prior assumption about $\overline{Y}$ closer to or farther from 0.5.

Let us redo these examples, but this time hold the margin of error d fixed at the original value of 0.045. The sample size when we assume $\overline{Y}$ is 0.35 is then

$$n = \frac{(1.959964)^2 \cdot 0.35 \cdot (1 - 0.35)}{0.045^2}$$

$$= 431.57131$$

while the sample size when we assume $\overline{Y}$ is 0.25 is

$$n = \frac{(1.959964)^2 \cdot 0.25 \cdot (1 - 0.25)}{0.045^2}$$

$$= 355.6904$$

Against the sample size for the case where we assume $\overline{Y}$ is 0.3 and the margin of error was 0.045 (i.e., $n = 398.37351$), this pattern makes far more sense in light of Figure 3.2. What is different in this paragraph, of course, is that we have varied only our assumption about $\overline{Y}$, and held d (and α) constant. That said, correctly guessing the magnitude of the change would still have been difficult.

Alternatively, consider one more variation. If we assume $\overline{Y} = 0.35$ and wish to have a margin of error d of 0.46, the sample size would be

$$n = \frac{(1.959964)^2 \cdot 0.35 \cdot (1 - 0.35)}{(0.046)^2}$$

$$= 413.01129$$

If we assume $\overline{Y} = 0.25$ but the margin of error d is 0.043, the sample size would be

$$n = \frac{(1.959964)^2 \cdot 0.25 \cdot (1 - 0.25)}{(0.043)^2}$$

$$= 389.54762$$

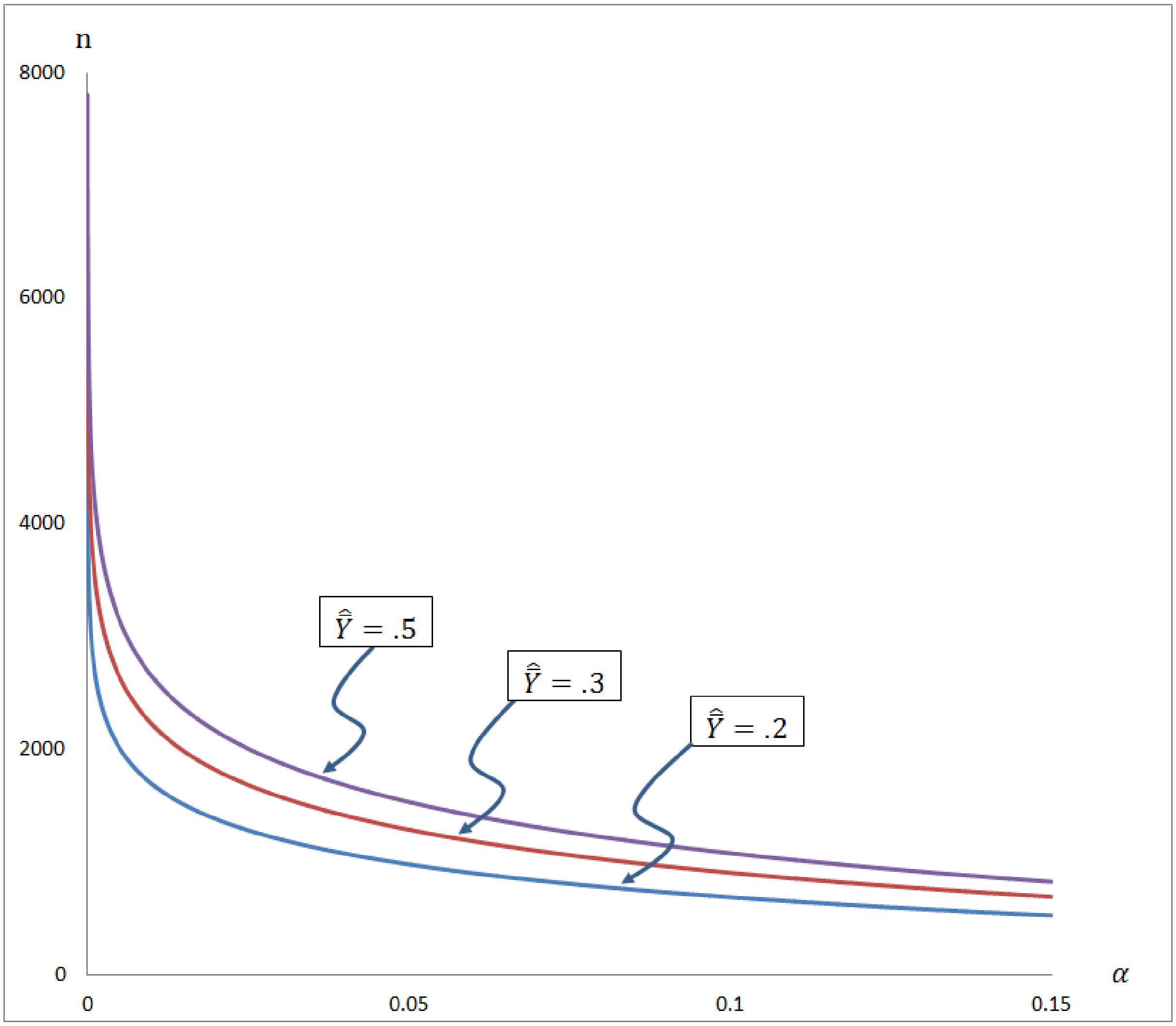


Figure 3.3. The influence of α on the sample size n

This still conforms to our pattern of expectations per Figure 3.2. So in this case, relative to the original example, the dynamics of $\hat{\overline{Y}} \cdot (1 - \hat{\overline{Y}})$ has overwhelmed the changes in d in determining the net change in sample size.

To round out the discussion of the potential consequences of varying the sampling parameters $\overline{Y}$ (or $\hat{\overline{Y}}$), α and d, in Figure 3.3 we consider the sample size n required for various values of α that might be considered in actual sample size estimation (ranging from 0.15 down to essentially zero[10]) for three prior assumptions for $\overline{Y}$ or $\hat{\overline{Y}}$ (0.2, 0.3 and 0.5) and $d = 0.025$. The patterns to the required sample sizes as we vary α illustrate the nonlinearity of the relationship between α and n. Moreover, the specifics of the pattern of nonlinearity for any given change in α clearly depend on $\overline{Y}$ (or $\hat{\overline{Y}}$).

What these examples and Figure 3.3 illustrate is the peril of trying to project the impact on required sample size of a proposed change in assumed $\overline{Y}$, d, or α short of actually performing the calculations. Even when we vary just one of these, the direction of the change might be clear but the magnitude is hard to predict. When several of these parameters are changed, even the direction of the change in required sample size is hard to predict. Even in the case of this simple precision motivated sample size estimation formula, nonlinearity introduces a great deal of uncertainty about the consequences of different values for $\overline{Y}$, d, and α, suggesting the need for caution about casually predicting those consequences.

We now briefly consider the case where the parameter of interest $\overline{Y}$ is continuous rather than a proportion. Proportions are intrinsically confined to take on values on the unit interval ($[0, 1]$) while "continuous" parameters, such as average body mass index, average family income, etc. are not confined in value between 0 and 1.[11] The derivation of a sample size estimate for a continuous parameter is exactly the same as for a proportion to this point:

$$d^2 = z^2_{1-\frac{\alpha}{2}} \cdot var\left(\hat{\overline{Y}}\right)$$

The only difference lies in how $var(\hat{\overline{Y}})$ is characterized.

Put simply, continuous variables have a different variance than proportions. The general formula for the variance in the continuous case is

$$var\left(\hat{\overline{Y}}\right) = \frac{\sigma^2_Y}{n}$$

[10] In practice, we estimated sample size down to an α value of $9.99999974737875 \cdot 10^{-6}$.

[11] Strictly speaking, a fully continuous variable or parameter is one that can take on any value from $(-\infty, \infty)$. Just because a variable or parameter is continuous does not mean that in practice it will ever actually take on the extreme values implied by this rather open interval. For instance, we drew 317,000,000 observations (to put this in some scale perspective, this is roughly one observation per resident of the United States) from a standard normal distribution (we did this in STATA with the seed set to 12345). Though a standard normal variable can theoretically take on any value on $(-\infty, \infty)$, the largest value drawn was 5.89725, while the minimum value drawn was -5.995363. Moreover, in practice, many of the "continuous" outcomes considered in health and the social sciences are restricted to be non-negative (that is, are restricted to $[0, \infty)$). For instance, neither average household income or body mass index can be negative. In this context, the value zero sometimes represents a legitimate corner solution and sometimes it reflects some kind of censoring. For the basic, and canonical, sample size estimation formulas that we consider here these subtle complications and distinctions are usually ignored.

where σ_Y is the standard deviation of the distribution of Y.[12] The sample size estimate is then based on

$$d^2 = z^2_{1-\frac{\alpha}{2}} \cdot \frac{\sigma^2_Y}{n}$$

Solving for n, we have

$$n = \frac{z^2_{1-\frac{\alpha}{2}} \cdot \sigma^2_Y}{d^2}$$

Note that this is very similar in structure to the sample size formula for a proportion. Of course, the variance of a proportion is a function of the level of the proportion, meaning that the main assumption that had to be made in the case of sample size for a proportion was just a likely value for $\hat{\overline{Y}}$. In the case of a continuous variable we will also need to make some assumptions regarding σ^2_Y. In the authors' experience, determining a reasonable (i.e., confidence-inspiring) prior for this has sometimes proven somewhat problematic given available prior evidence.

Nonetheless, let us consider a numerical example to demonstrate how to perform the sample size estimation. Suppose, for instance, that we wish to estimate the average body mass index (defined by the body weight in kilograms divided by the squared height in meters) among adults ages 20 and older. Suppose that we have priors for $\overline{Y}$ of 23.0 and σ^2_Y of 25, respectively, and we wish to have an error d of 0.5 at the significance level of 0.05.

The sample size estimate is then

$$n = \frac{(1.959964)^2 \cdot 25}{(0.5)^2}$$

$$= 384.14588 \approx 385$$

We would need a sample of 385 adults to estimate the average body mass index.

We have sometimes seen a t-statistic from Student's t-distribution applied in sample size estimation along the lines of the present discussion. In formal terms, the t-distribution is **leptokurtic**, a fancy way of saying that a distribution has more weight in the tails than the normal distribution. This becomes less true as the degrees of freedom (essentially, the sample size minus the number of parameters to be estimated, which in this case is $n-1$) increases, until eventually the t-distribution becomes essentially indistinguishable from the normal distribution. Generally, except at very small sample sizes, the difference between the Student's t and standard normal distributions is not that large. By roughly 30 degrees of freedom the two distributions are virtually identical. Nonetheless, the use of the t-statistic for sample size estimation is somewhat conceptually strange since its shape (and hence critical values) depends on the degrees of freedom which is in turn a function of n (the

[12] There is often confusion about the distinction between the standard deviation σ_Y and the standard error

$$\frac{\sigma_Y}{\sqrt{n}}$$

The standard deviation is a measure of the intrinsic variability of a random variable. It tells us something about the variation in the value of that variable around its mean as observations are generated per the process behind that variable (such as normality). The standard error is a sample feature that tells us something about the precision of an estimate of a sample mean.

very thing we are trying to estimate in the sample size estimation process). If for no other reason than this, we tend to rely on z-statistics from the standard normal distribution for sample size estimation in cases where precision is the main consideration.

STATA Output 3.1

```
Estimated sample size for a one-sample proportion test
Score z test
Ho: p = p0  versus  Ha: p != p0

Study parameters:

        alpha =    0.0500
        power =    0.5000
        delta =    0.0450
           p0 =    0.3000
           pa =    0.3450

Estimated sample size:

            N =       399
```

In STATA, the sample size estimations for a specified precision in estimating the population average $\overline{Y}$ can be performed by using the command `power`. We demonstrate how to use the STATA command `power` for estimating a sample size for a specified precision around two types of population average $\overline{Y}$: first when $\overline{Y}$ represents a proportion, and next when $\overline{Y}$ represents an average of a continuous variable Y. We employ the examples we used in our manual calculation for estimating sample size and confirm the results in STATA.

First, recall that when $\overline{Y}$ represents a population proportion, the sample size estimator is defined by

$$n = \frac{z^2_{1-\frac{\alpha}{2}} \cdot \hat{\overline{Y}} \cdot \left(1 - \hat{\overline{Y}}\right)}{d^2}$$

The basic syntax of the command `power` for estimating sample size is

```
power oneproportion p0 pa [, power(numlist) options]
```

where `oneproportion` stands for a one-sample proportion test, `p0` denotes Y^*, and `pa` denotes $Y^* + d$.[13] A user specifies the desired level of power (1-β) in `power(numlist)`, and other options, including the desired significance level (α), in `options`.[14]

Recall that our example with a prior for $\overline{Y}$ of 0.3 and d of 0.045 (15% of the true population parameter value) at the significance level of 0.05 resulted in the estimated sample size of 399 in our manual calculation. To perform the sample size estimation in STATA, we feed the `power` command

[13] Alternatively, it can be specified as:

```
power oneproportion p0, diff(numlist) power(numlist)
```

where `diff(numlist)` denotes d.

[14] For the entire list of available methods and options in `power`, refer to the STATA user's manual. Also, one can have a quick review by typing `help power` in the STATA command window.

as:

```
power oneproportion 0.3 0.345, power(0.5) alpha(0.05)
```

which produces results shown in STATA Output 3.1. Notice that the statistical power is specified at 0.5 (50%), which is equivalent to $\Pr[Z > z]$ when $z = 0$, so that it does not play a role in the sample size estimation when precision is our motivation (which does not involve hypothesis testing).[15] We will define the statistical power and its role in the next section when we discuss hypothesis testing.

Under the first line `Estimated sample size for a one-sample proportion test`, we see `Score z test`, which confirms that a Z-statistic is employed for the sample size estimation. Notice that, while null and alternative hypotheses are specified in `Ho: p = p0 versus Ha: p != p0`, our sample size estimation is driven by precision and not by hypothesis testing because we have specified the statistical power at 0.5. Under `Study parameters` we see the specified values of our parameters, namely α in `alpha`, $1 - \beta$ in `power`, d in `delta`, Y^* in `p0`, and $Y^* + d$ in `pa`. Finally, the estimated sample size is provided in the last output `N =399`, which confirms our manual calculation.

Next, we demonstrate how to use the STATA program `power` when $\overline{Y}$ represents a population average of a continuous Y. Recall that when $\overline{Y}$ represents a population average, the sample size estimator is defined by

$$n = \frac{z^2_{1-\frac{\alpha}{2}} \cdot \sigma^2_Y}{d^2}$$

The basic syntax of the command `power` for performing the sample size estimation is:

```
power onemean m0 ma [, power(numlist) options]
```

where `onemean` stands for a one-sample mean test, `m0` denotes Y^*, and `ma` denotes $Y^* + d$.[16] Again, a user specifies the desired level of power (1-β) in `power(numlist)`, and other options, including the desired significance level (α), and the standard deviation (σ_Y) in `options`.

Recall that our example with a prior for $\overline{Y}$ of 23, d of 0.5, and σ^2_Y of 25 at the significance level of 0.05 resulted in the estimated sample size of 385 in our manual calculation. To perform the sample size estimation in STATA, we feed the `power` command as:

```
power onemean 23 23.5, alpha(0.05) power(0.5) sd(5) knownsd
```

which produces STATA Output 3.2. Notice that the statistical power is again specified at 0.5 (50%), which is equivalent to $\Pr[Z > z]$ when $z = 0$. Also note that we specify σ_Y instead of σ^2_Y in `sd(numlist)`, along with `knownsd` to indicate that σ_Y is assumed to be known.[17]

[15] The STATA `power` command is built for multiple purposes, including power and sample size estimations for precision and hypothesis testing of proportions and means. The sample size estimation driven by precision (instead of hypothesis testing) can be seen as a special case of a sample size estimation based on hypothesis testing but with the statistical power of 0.5 (50%). Therefore for using the STATA `power` command, we estimate sample size driven by precision within the framework of sample size estimation driven by hypothesis testing, but we should specify the power at 0.5.

[16] Alternatively, it can be specified as:

```
power onemean m0, diff(numlist) power(numlist)
```

where `diff(numlist)` denotes d.

[17] The default in the STATA command `power` provides a sample size estimation assuming that the standard deviation is unknown, which results in a larger sample size estimate, to account for the uncertainty associated with the prior of the standard deviation.

STATA Output 3.2

```
Estimated sample size for a one-sample mean test
z test
Ho: m = m0  versus  Ha: m != m0

Study parameters:

        alpha =    0.0500
        power =    0.5000
        delta =    0.1000
           m0 =   23.0000
           ma =   23.5000
           sd =    5.0000

Estimated sample size:

            N =       385
```

Under the first line `Estimated sample size for a one-sample mean test`, we see `z test` which confirms that a Z-statistic is employed for the sample size estimation. Again, while null and alternative hypotheses are specified in `Ho: m = m0 versus Ha: m != m0`, our sample size estimation is driven by precision and not by hypothesis testing because of the specification of power at 0.5. Under `Study parameters` we see the specified values of our parameters, namely α in `alpha`, $1-\beta$ in `power`, d in `delta`, Y^* in `m0`, Y^*+d in `ma`, and σ_Y in `sd`. Finally, the estimated sample size is provided in the last output $N = 385$, which confirms our manual calculation.

3.1.2 Basic Hypothesis Testing

Until now, the goal in sample size estimation has been to find the sample size that likely yields a desirable degree of precision to our estimate $\hat{\overline{Y}}$ of the parameter $\overline{Y}$. We weren't seeking explicitly to *test* anything about $\overline{Y}$ through our estimate $\hat{\overline{Y}}$ (or, as we will see, through estimate*s*).

Testing is more central to the story of impact evaluation than precision alone. We typically want to know whether a program made a difference, and, in the most foundational sense, we assess this by testing whether there is a difference in the parameter of interest $\overline{Y}$ between the participant and control/comparison populations.

For the present purposes, we avoid much explicit modelling or notation specific to impact evaluation (such as conventional "potential outcome" notation) to instead focus on the general question of how to estimate sample size for testing hypotheses. We begin with a few broad concepts. To begin with, we wish to test whether something is true of a population parameter (later we will expand this to consider the parameters of two different populations). This truth about the population parameter is called the null hypothesis, which is usually denoted H_0. For instance, if the population parameter of interest is $\overline{Y}$, the null hypothesis is typically something along the lines of

$$H_0 : \overline{Y} = Y^*$$

The null hypothesis in this case is that the population parameter $\overline{Y}$ equals some constant Y^*.

So stated, the null hypothesis H_0 naturally begs the question "as opposed to what?" That alternative possibility for the null hypothesis is referred to as the alternative hypothesis and typically indicated by H_a. One possible "general" alternative hypothesis might be

$$H_a : \overline{Y} > Y^*$$

We would reject the null hypothesis if $\overline{Y} > Y^*$ (or, more precisely, if our estimates suggested that this is the case).

While this "general" alternative hypothesis is theoretically persuasive, testing it (so stated) in practice with a sample from a population of interest might be meaningless. The reason is that *even if the null were true* we would expect our estimate would sometimes be greater than Y^* simply due to ordinary sampling variation. Therefore, in practice we will need to frame the alternative hypothesis in terms of a more specific threshold for a violation $\overline{Y} > Y^*$ that we wish to detect.

An example of a more specific alternative hypothesis might be

$$H_a : \overline{Y} \geq Y^* + d$$

where $d > 0$.[18] This is what is called a "one-sided" alternative (because it involves an alternative possibility on only one side of the null hypothesis).

A "two-sided" alternative would be something along the lines of

$$H_a : \overline{Y} \neq Y^*$$

Obviously, this is once again a rather broad alternative in the sense that it could accommodate many alternative (to Y^*) values for $\overline{Y}$, many of which we could expect to see in actual estimates due to routine sampling variation.[19] Once again we will need to consider a more specific alternative.

[18] We could have also considered the general null hypothesis

$$H_a : \overline{Y} < Y^*$$

leading to the more specific null

$$H_a : \overline{Y} \leq Y^* - d$$

for some $d > 0$. Or some $d < 0$. Or whatever. The discussion will proceed with H_a as framed in the main text without loss of generality in terms of the logic of how sample size estimation is performed. We do it simply to make the discussion a bit conceptually simpler for the reader.

[19] For instance, suppose that

$$H_0 : \overline{Y} = 12$$

As we will see, in reasonable-sized empirical samples it would typically be impossible to test this against an alternative value of, say, 12.00001, with any real degree of faith in our results (where "faith" is a concept we will shortly make more concrete).

We will typically set up two-sided tests with two possible alternatives in mind. These are

$$H_a : \overline{Y} \geq Y^* + d$$

and

$$H_a : \overline{Y} \leq Y^* - d$$

In other words, whether it is for the one- or two-sided test, we are usually testing whether the parameter is more than d away from the null hypothesis value.

The plan in either the one- or two-sided case is to test the null hypothesis H_0 against the alternative H_a. The test statistic will generally be formed from estimates and will not, therefore, reveal with certainty the population truth. Another way of saying this is that the test statistic can lead us to the wrong conclusion. Two such errors are possible:

1. Rejecting a True Null Hypotheses H_0: This is referred to as a "Type-I" error. As we will see shortly, our testing approach is set up to assign this error a probability α of happening.

2. Rejecting a True Alternative Hypothesis H_a: This is referred to as a "Type-II" error. Our testing approach is set up to assign this error a probability β of happening.

These two errors are a recurrent theme in sample size estimation because they are the two obvious testing errors that can occur under the departing framework of testing a null hypothesis against some alternative. The probability of a Type-I error, α, is usually referred to as the significance of a test. It is also sometimes called the *size* of the test. One minus the probability of a Type-II error (i.e., $1 - \beta$) is usually referred to as the "power" of a test.

It is worth pausing to think about the importance of these errors in the context of impact evaluation. Typically, in impact evaluation we want to know whether a program made a difference. The usual approach is a null hypothesis that the program made no difference. The alternative hypothesis is that the program made a difference to a degree of d or more. Given that a Type-II error thus might lead to the cancellation of a program that was actually working (and, frankly, given that there is often an unavoidable human emotional stake in believing that a programmatic investment was worth it) the focus is often on Type-II errors and the power of a test. In short, in practice folks seem particularly concerned with committing a Type-II error.

As a result, one commonly hears language in impact evaluation practice such as "our sample was powered on" such and such indicator. For instance, one might hear "our sample was powered on detecting 5 percent change in the contraceptive prevalence rate for 15–49 year olds." What the speaker presumably means by this is that the sample size was estimated based on the goal of detecting with a certain degree of power a 5 percentage-point change.

This language is truly misleading: there are <u>two</u> types of testing errors that can occur, and sample size estimation is driven by the desire to keep the probability of either of them to tolerable levels. Choices about the probabilities of both Type-I and Type-II errors matter for sample size. For instance, all other things being equal, changing the probability of a Type-I error (α) from 1 to 10 percent often roughly halves the required sample size in the tests we consider in this subsection.

It is also probably fairly inappropriate on a philosophical level as well as Type-I errors are also programmatically important. There is, in the face of scarce resources, a real cost associated with continuing to plow resources into a program that isn't working. The resources dedicated to ineffective programs could have been committed to some kind of alternative programming that might have actually worked. Given this, we would argue that there should be more thought given to the significance level of a test.

As matters stand, the significance level is often essentially completely overlooked in sample size estimation. Instead, the default convention of practice is, more often than not, simply to adopt a significance level of 5 percent, or $\alpha = 0.05$. The inertia behind this default choice is unfortunate: the probability of a Type-I error should be as important a consideration in sample size estimation as the probability of a Type-II error.

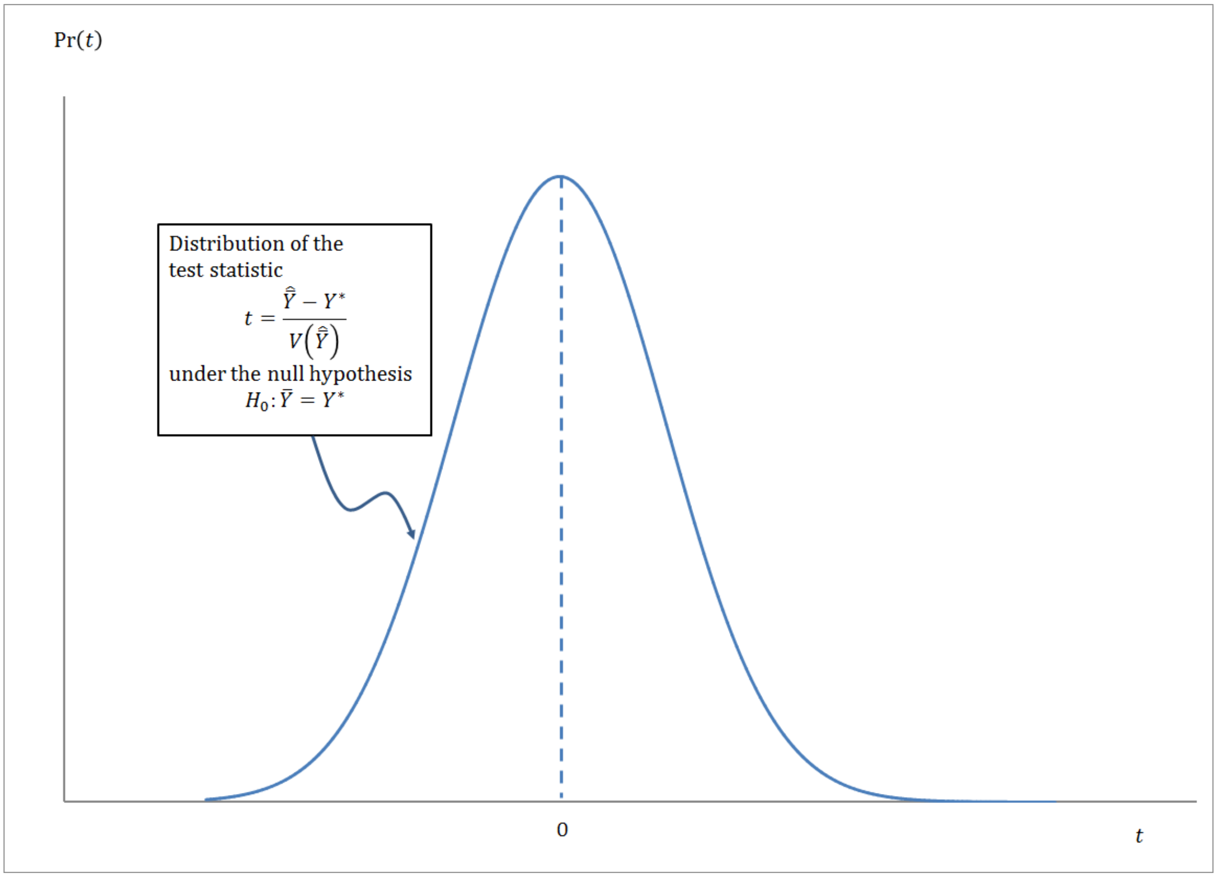


Figure 3.4. The distribution of the test statistic under the null

Let us begin our technical discussion with the case of determining sample size for a single sample when we wish to test whether the population parameter equals some value. The starting point is the null hypothesis

$$H_0 : \overline{Y} = Y^*$$

The basis for developing sample size estimators is the basic test statistic

$$\frac{\hat{\overline{Y}} - Y^*}{\sqrt{V\left(\hat{\overline{Y}}\right)}}$$

where $\hat{\overline{Y}}$ is an estimate of $\overline{Y}$ and $V(\hat{\overline{Y}})$ is the sampling variance of that estimate. Assuming that $\hat{\overline{Y}}$ emerges from an unbiased sample selection and estimation process, under the null hypothesis the expectation of the estimate is

$$E\left(\hat{\overline{Y}}\right) = Y^*$$

The test statistic should thus, under the null hypothesis, be centered on 0.

The distribution of the test statistic

$$t = \frac{\hat{\overline{Y}} - Y^*}{\sqrt{V\left(\hat{\overline{Y}}\right)}}$$

under the null hypothesis is shown in Figure 3.4. At this point the "*t*" is just meant to indicate "test statistic" and the reader should refrain from the temptation to assume that this means Student's t-distribution.[20] Because we will later rely on a symmetric, unimodal distribution (that is, a distribution with one peak at e,[21] and the same mass for every interval $[e - q, e + q]$) the shape in Figure 3.4 roughly conforms to what we will be expecting for the assumed distribution of our test statistic t.

The distribution of t peaks at 0 in Figure 3.4 because, as graphed, it assumes that the null hypothesis

$$H_0 : \overline{Y} = Y^*$$

is true. If that was the case then we would expect the test statistic

$$t = \frac{\hat{\overline{Y}} - Y^*}{\sqrt{V\left(\hat{\overline{Y}}\right)}}$$

to equal 0, meaning that its expectation or average across many, many samples should be 0. We expect this essentially because $\hat{\overline{Y}}$ is expected to equal Y^* (i.e., its expectation or average across many, many samples should be Y^*) and hence $\hat{\overline{Y}} - Y^*$ has an expected value of 0. Put slightly differently, this is the distribution of the test statistic under one of the hypotheses (in this case the null $H_0 : \overline{Y} = Y^*$) and all of the "action" for understanding the central tendency of the statistic

[20] Indeed, we only introduce the label t to the statistic in the interest of tidy axis labelling in Figure 3.4 and the figures to follow and cleaner looking probability statements in the paragraphs to follow. The use of a generic "t" for the purpose of sample size estimation discussion is not without precedent (for example, Valliant, et al., 2013) follow the same convention). Some prefer to use "z" at an analogous stage of their discussion (such as Fleiss, Levin, & Paik, 2003). We follow the convention "t" as in "test statistic" because we have not made clear the specific distribution we will use to operationalize the sample size estimation. Furthermore, we flipped a coin and "z" lost.

[21] In Figure 3.4 $e = 0$.

under the assumed hypothesis is in the numerator of the statistic.[22]

We call it the test statistic because we will use it to test whether our sample seems to support the null hypothesis. Specifically, we decide on the probability of a Type-I error that we are willing to tolerate and then accept the null if the test statistic value seems consistent with that null hypothesis. There are three things that need to be done at this point to make that a viable test.

First, we must describe the alternative hypothesis a bit. In the current example, let us consider a one-sided test with the alternative hypothesis[23]

$$H_a : \overline{Y} > Y^*$$

We will accept the null hypothesis if the test is consistent with the null. We will reject if the test statistic t is too large in value.

Second, we need to decide on our tolerance for a Type-I error by deciding on a level of significance (or, alternatively, "test size"). In other words, we need to decide on the probability of a Type-I error occurring. A typical choice for the level of significance is $\alpha = 0.05$, but for now we simply assume that some generic choice α was made for the probability of the Type-I error.

Finally, we must turn to some distribution to determine a critical value for the test statistic. Having set the level of significance to α, the critical value is, given our null hypothesis, that value for the test statistic above which the probability mass (or the total probability of experiencing some value from that range) is α. In other words, it is the value for the rejection (of the null hypothesis) threshold for the test statistic such that the probability of rejecting the null incorrectly (i.e., of a Type-I error) is no more than α. The concept of the critical value for a one-sided test (the type we are now considering) is illustrated in Figure 3.5 by the threshold c.

Assigning a specific value to c given our decision about α requires making a distributional assumption: we need to make some assumption about the distribution of the test statistic under the null hypothesis. Given the form of the test statistic, the two obvious candidates for a distribution from which to find critical values are the standard normal distribution and Student's t-distribution. At first glance, the t-distribution might seem more appropriate since the eventual test statistic

$$t = \frac{\hat{\overline{Y}} - Y^*}{\sqrt{V\left(\hat{\overline{Y}}\right)}}$$

will likely rely on an *estimate* of the variance

$$\hat{V}\left(\hat{\overline{Y}}\right)$$

[22] The denominator is a variance which essentially by definition must exceed zero in value.

[23] Without loss of generality we could have considered the alternative hypothesis

$$H_a : \overline{Y} < Y^*$$

in the denominator. However, the convention in sample size estimation of this sort is usually to use the standard normal distribution. There are likely three reasons for this. First, as we mentioned previously, as the number of degrees of freedom (in this case the sample size n minus 1) grows, the differences between the standard normal and t-distributions become less and less apparent (and this happens very quickly: By perhaps 30 or 40 degrees of freedom the differences between the two distributions, and hence the critical values they would suggest, are quite small). Second, because the t-distribution's shape depends on the degrees of freedom, which is a function of a sample size, it is a somewhat awkward fit for the task of sample size estimation. Its critical values for significance level α will technically depend on the very thing we are trying to estimate, the sample size n. Finally, sample size estimation for proportions tends to depend on a normality assumption (more on this below).

Once we have settled all of these matters we can determine the critical value for significance level α for the standard normal distribution. We reject the null hypothesis if

$$t > c = z_{1-\alpha}$$

The rejection region is illustrated by the shaded area in Figure 3.5. The shaded area under the

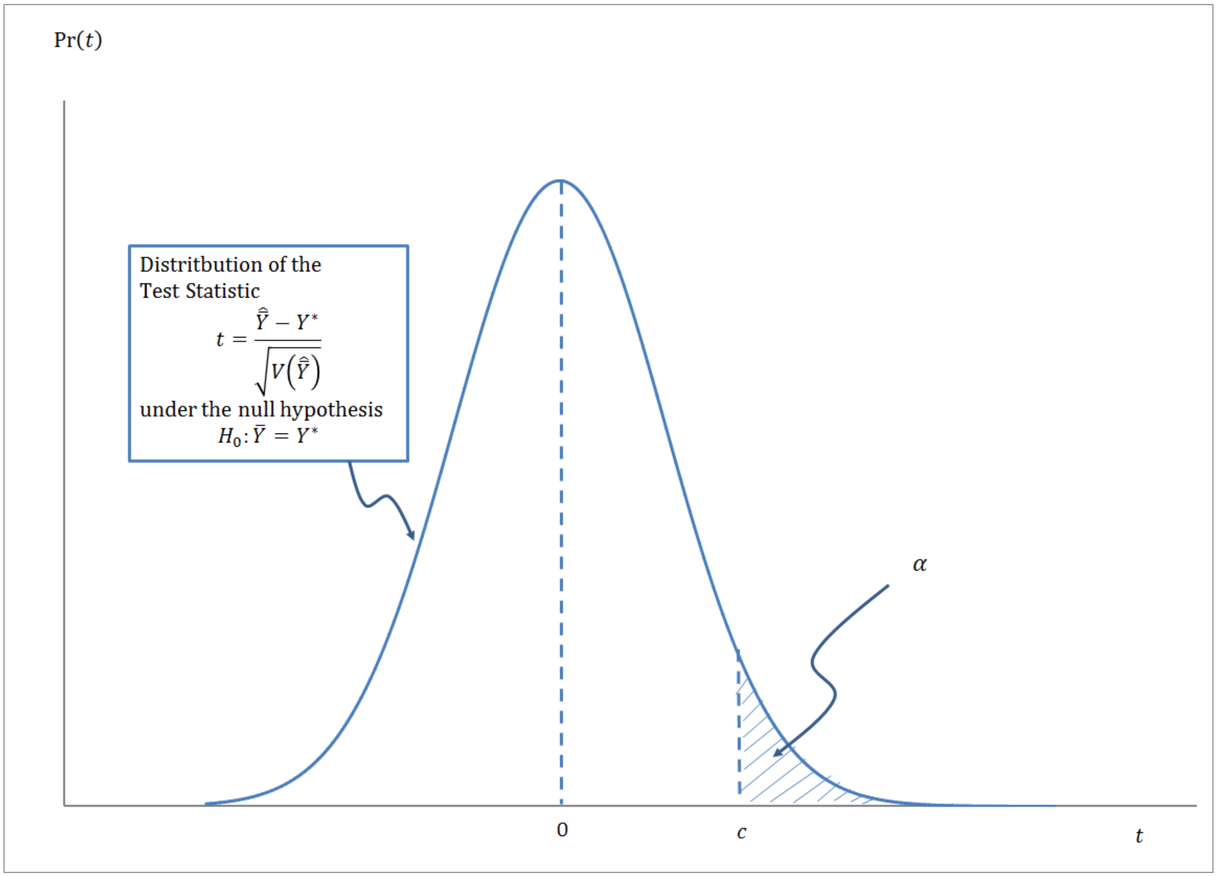

Pr(t)
Distritbution of the Test Statistic
$t = \frac{\hat{\bar{Y}} - Y^*}{\sqrt{V(\hat{\bar{Y}})}}$
under the null hypothesis
$H_0: \bar{Y} = Y^*$
α
0
c
t

Figure 3.5. The critical value for significance level α

curve has an area of α, the significance level of the test. In terms of the specific values for the critical value, it depends on one's choice for the significance level α. In Table 3.1, we list some critical values associated with commonly adopted significance levels (with $\alpha = 0.05$ likely being by far the most common choice).

Table 3.1. Critical values

α	**Critical Value** $c = Z_{1-\alpha}$
0.2	0.8416212
0.15	1.0364334
0.1	1.2815516
0.05	1.6448536
0.025	1.9599640
0.01	2.3263479

Notice that as the level of significance (i.e., the probability of a Type-I error) falls, the critical value grows. The intuitive reason for this is that to be more sure that we are not making a Type-I error, we need to allow for the possibility of experiencing a greater range of values for the test statistic t under the null hypothesis H_0. In other words, if we want to make a Type-I error less likely,

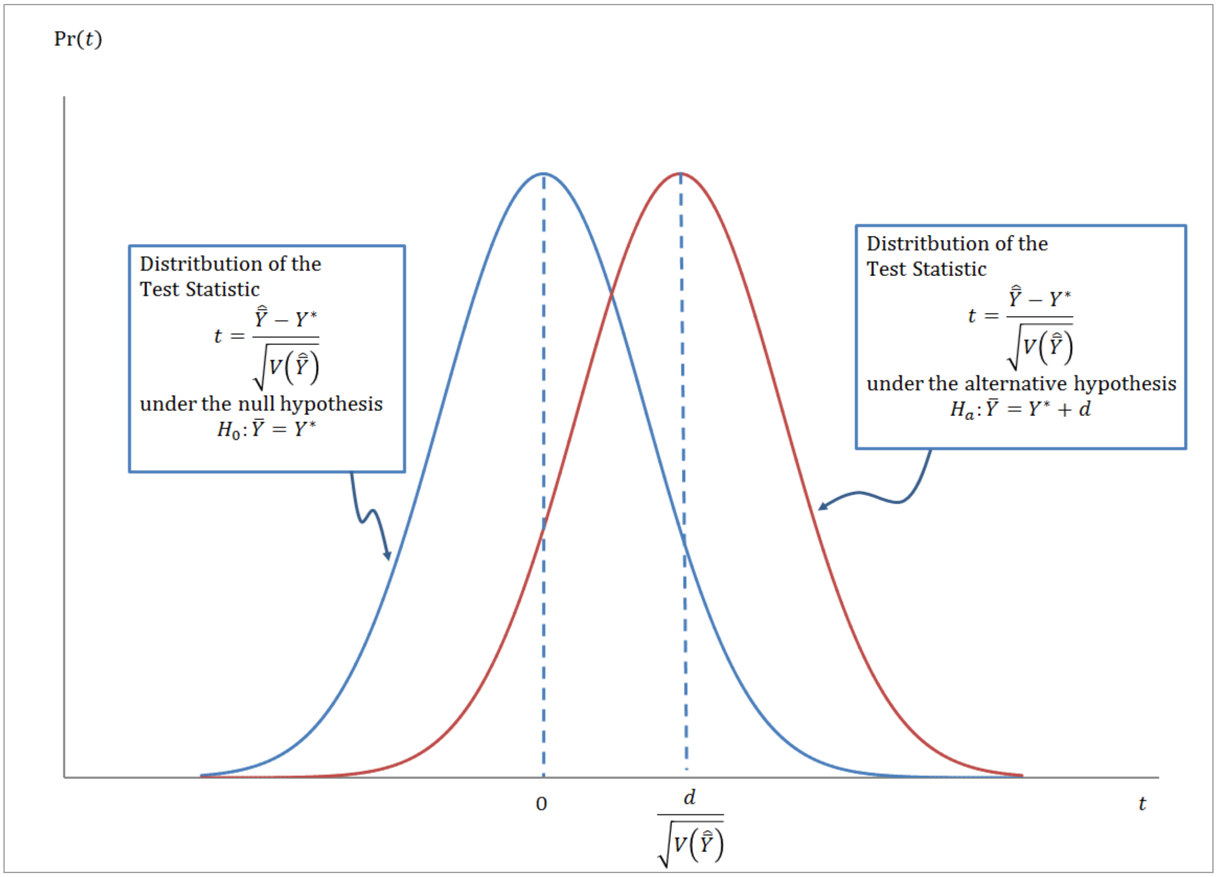


Figure 3.6. The distribution of the test statistic under the null and alternative hypotheses

we have to adopt the idea of rejecting the null only under increasingly strong evidence (where the larger the value of the test statistic t the less consistent is the test statistic with the null hypothesis).

So far we have made little use of the alternative hypothesis

$$H_a : \overline{Y} > Y^*$$

Its only real role so far has been to shape our thinking about the rejection criteria for our test statistic. In particular, since it involved a one-sided inequality it was apparent that we were confronted with a one-sided test.

However, for sample size estimation, the specifics of the null hypothesis are in fact quite important. It is not enough to say

$$H_a : \overline{Y} > Y^*$$

We need to say how much difference from the null is important to us. Suppose, for instance, that Y^* is 12. It would likely require an enormous sample size to detect a difference between 12 and 12.0000001 (and how important would it be to detect such a minuscule difference anyway?).

To move to the next stage in the development of a sample size estimation formula, we need to articulate how much $\overline{Y}$ needs to exceed Y^* before we consider it important enough to detect with

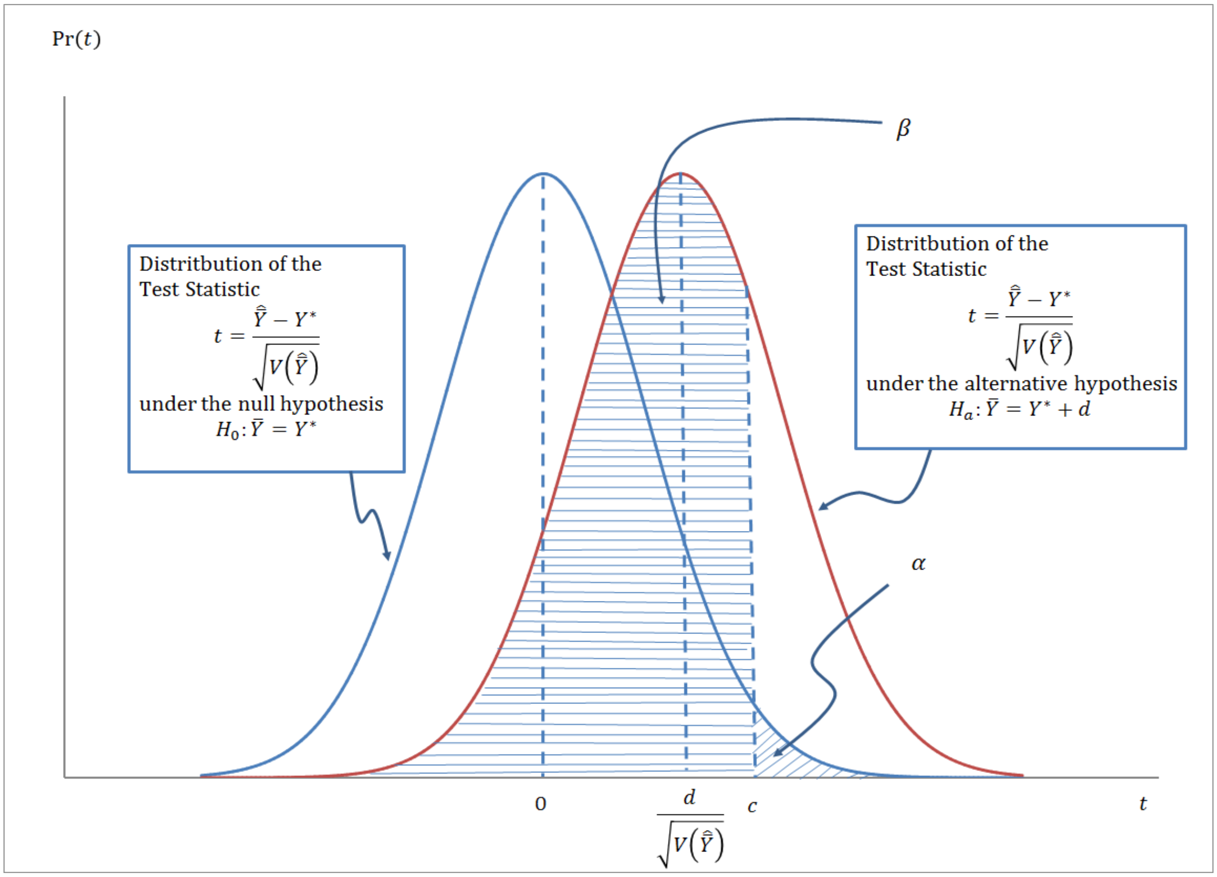


Figure 3.7. The probabilities of Type-I (α) and Type-II (β) errors

our desired degree of power. This is in the end a question driven by the objectives of the survey. For now, we decide that we wish to be able to detect a generic difference d (where $d > 0$) from Y^*. The alternative hypothesis is

$$H_a : \overline{Y} \geq Y^* + d$$

What we are saying with an alternative hypothesis of this form is that we want a sample size large enough to detect a difference in the true parameter from the value under the null hypothesis (of Y^*) of d or more (in one direction under this alternative). The obvious error in this case would be to fail to reject the null hypothesis when the alternative (that the true population value of the parameter $\overline{Y}$ is greater than or equal to $Y^* + d$) is actually true. This would be a Type-II error and its probability would be β, making the power of the test $1 - \beta$.

As stated, the alternative hypothesis accommodates an infinite number of possible true values for the population parameter: It will hold as long as $\overline{Y} \geq Y^* + d$. So in other words, it encompasses any possible population value greater than or equal to $Y^* + d$. Our sample size estimate will be large enough to provide a sample size of sufficient size to accept[24] the alternative hypothesis if the true population parameter is $Y^* + d$ or larger with probability $1 - \beta$.

Because of the multitude of true population parameter values consistent with the alternative hypothesis, it is most straightforward to illustrate power for the case of

$$H_a : \overline{Y} = Y^* + d$$

If this alternative hypothesis is true, the test statistic

$$t = \frac{\hat{\overline{Y}} - Y^*}{\sqrt{V\left(\hat{\overline{Y}}\right)}}$$

has a distribution centered on $\frac{d}{\sqrt{V\left(\hat{\overline{Y}}\right)}}$, as shown in Figure 3.6. The distribution of the test statistic under this alternative hypothesis is centered on d because we assume that the expectation of $\hat{\overline{Y}}$ is $Y^* + d$, and the expectation of $\hat{\overline{Y}} - Y^*$ is d. The probability of a Type-II error, β, is then the area under the distribution of the test statistic under the null hypothesis where we would accept the null hypothesis (in this case values for the test statistic less than the critical value $c = Z_{1-\alpha}$). This area equal in size to β, the probability of a Type-II error, is shown in Figure 3.7.[25]

Returning to our test statistic

$$t = \frac{\hat{\overline{Y}} - Y^*}{\sqrt{V\left(\hat{\overline{Y}}\right)}}$$

the power (i.e., one minus the probability of a Type-II error) can be written (eschewing c to

[24] Some would say "fail to reject."

[25] Figures 3.6 and 3.7 were drawn for visual clarity. In an actual application, this would represent very low power $1 - \beta$.

represent the critical value and instead just writing $z_{1-\alpha}$)

$$1-\beta = Pr\left(t > z_{1-\alpha} | H_a\right) = Pr\left(\frac{\hat{\overline{Y}} - Y^*}{\sqrt{V\left(\hat{\overline{Y}}\right)}} > z_{1-\alpha} \middle| \overline{Y} = Y^* + d\right)$$

The next step involves a small trick — subtracting

$$\frac{d}{\sqrt{V\left(\hat{\overline{Y}}\right)}}$$

from both sides of the inequality at the center of probability of rejecting the null if H_a is true

$$Pr\left(\frac{\hat{\overline{Y}} - Y^*}{\sqrt{V\left(\hat{\overline{Y}}\right)}} - \frac{d}{\sqrt{V\left(\hat{\overline{Y}}\right)}} > z_{1-\alpha} - \frac{d}{\sqrt{V\left(\hat{\overline{Y}}\right)}} \middle| \overline{Y} = Y^* + d\right)$$

Next we assume that

$$\frac{\hat{\overline{Y}} - Y^*}{\sqrt{V\left(\hat{\overline{Y}}\right)}} - \frac{d}{\sqrt{V\left(\hat{\overline{Y}}\right)}}$$

follows a standard normal distribution (which is reasonable given both its structure and our approach to quantities of that structure thus far). In other words,

$$Pr\left(\frac{\hat{\overline{Y}} - Y^*}{\sqrt{V\left(\hat{\overline{Y}}\right)}} - \frac{d}{\sqrt{V\left(\hat{\overline{Y}}\right)}} > z_{1-\alpha} - \frac{d}{\sqrt{V\left(\hat{\overline{Y}}\right)}} \middle| \overline{Y} = Y^* + d\right)$$

$$= Pr\left(Z > z_{1-\alpha} - \frac{d}{\sqrt{V\left(\hat{\overline{Y}}\right)}} \middle| \overline{Y} = Y^* + d\right)$$

where Z means a standard normal distribution. Notice that for any given values of

$$z_{1-\alpha}$$

and

$$\frac{d}{\sqrt{V\left(\hat{\overline{Y}}\right)}}$$

we can calculate the value of this probability using the standard normal distribution. This is the "power function" of the text.

We want to achieve a power of $1-\beta$. In other words, we want

$$Pr\left(Z > z_{1-\alpha} - \frac{d}{\sqrt{V\left(\hat{\overline{Y}}\right)}}\,\middle|\, \overline{Y} = Y^* + d\right) = 1-\beta$$

But since Z is a standard normal random variable, this means that we need

$$Pr\left(Z > -z_{1-\beta} \,\middle|\, \overline{Y} = Y^* + d\right) = 1-\beta$$

Therefore, we need

$$-z_{1-\beta} = z_{1-\alpha} - \frac{d}{\sqrt{V\left(\hat{\overline{Y}}\right)}}$$

Rearranging a bit, we have

$$z_{1-\alpha} + z_{1-\beta} = \frac{d}{\sqrt{V\left(\hat{\overline{Y}}\right)}}$$

then

$$\left(\frac{d}{z_{1-\alpha} + z_{1-\beta}}\right)^2 = V\left(\hat{\overline{Y}}\right)$$

We are now almost done developing a sample size estimate.

What remains is to deal with the variance of the estimate

$$V\left(\hat{\overline{Y}}\right)$$

If Y is a continuous variable, we can think of this as equalling

$$V\left(\hat{\overline{Y}}\right) = \frac{\sigma_Y^2}{n}$$

We then have

$$\left(\frac{d}{z_{1-\alpha} + z_{1-\beta}}\right)^2 = \frac{\sigma_Y^2}{n}$$

Solving for n, we have

$$n = \frac{\left(\sigma_Y \cdot \left(z_{1-\alpha} + z_{1-\beta}\right)\right)^2}{d^2}$$

This sample size is increasing as power $1-\beta$ increases, decreasing as significance α falls (it is not just about power that matters for sample size!) and increasing as the difference d between the null and the alternative that we wish to select falls. This makes sense: We need more information to be more sure that we aren't committing a Type-I or -II error and to be able to differentiate the null hypothesis H_0 from closer alternatives.

This sample size estimation can be performed using the STATA command `power` by specifying:

```
   power onemean m0 ma [, alpha(numlist) power(numlist) sd(numlist)
knownsd onesided]
```

where `onemean` stands for a one-sample mean test, and `m0` and `ma` denote Y^* and $Y^* + d$, respectively. In the options, we specify the significance level (α) in `alpha(numlist)`, the power level $(1-\beta)$in `power(numlist)`, and the standard deviation (σ_Y) in `sd(numlist)`. The sample size estimation formula is based on the assumption that the standard deviation is known, so we specify `knownsd`. Finally, we indicate that the test is one-sided by specifying `onesided`.[26]

We next consider the case where Y is a binary variable and $\overline{Y}$ represents a proportion. This is a little bit more complicated for two reasons. First, and less importantly, we have to assume a normal distribution for our test statistic. A proportion is rooted in a discrete binary variable which cannot follow the normal distribution and hence cannot follow the t-distribution (the t-distribution is premised on the normality of the underlying variable[s] behind it). Nonetheless, our test statistic is roughly normal (particularly in bigger samples) and so we tend to rely on normality.

The bigger problem is that the variance of the estimate of a proportion depends on its value in a way that is not the case with a continuous variable. The variance of a proportion $\overline{Y}$ is given by

$$\frac{\overline{Y} \cdot \left(1 - \overline{Y}\right)}{n}$$

The variance of the estimate if the null hypothesis value is Y^* is hence

$$\frac{Y^* \cdot (1 - Y^*)}{n}$$

To simplify things, we write the alternative as

$$H_a : \overline{Y} \geq Y^* + d = Y_a$$

The variance if the specific alternative

$$H_a : \overline{Y} = Y_a$$

was true would then be

$$\frac{Y_a \cdot (1 - Y_a)}{n}$$

The sticky wicket for sample size estimation with a proportion is thus that we cannot rely on one single value for $V(\hat{\overline{Y}})$.

[26] Alternatively, the same sample size estimation can be performed by specifying:

```
  power onemean m0[, diff(numlist) alpha(numlist) power(numlist) sd(numlist) knownsd
onesided]
```

where `diff(numlist)` denotes d.

To get to a sample size, let us back track to

$$Pr\left(t > z_{1-\alpha} | H_a\right) = Pr\left(\frac{\hat{\overline{Y}} - Y^*}{\sqrt{V\left(\hat{\overline{Y}}\right)}} > z_{1-\alpha} \,\middle|\, \overline{Y} = Y^* + d\right)$$

So the issue is thus that we need some prior value at which to peg the variance of the estimate under the null hypothesis. As we have mentioned, if the null hypothesis $H_0 : \overline{Y} = Y^*$ is true, the sampling variance would be

$$\frac{Y^* \cdot (1 - Y^*)}{n}$$

It helps at this point to remember a bit the logic of testing. We craft the test statistic

$$t = \frac{\hat{\overline{Y}} - Y^*}{\sqrt{V\left(\hat{\overline{Y}}\right)}}$$

under the assumption that the null hypothesis is true. We reject if $t > c = z_{1-\alpha}$. In other words, we reject if the test statistic seems unlikely under the assumption that the null is true. The point, however, is that the statistic is based on the assumption that the null hypothesis is true.

The most sensible way to proceed is thus simply to use the variance when the proportion equals the null value:

$$V\left(\hat{\overline{Y}}\right) = \frac{Y^* \cdot (1 - Y^*)}{n}$$

We then have

$$\begin{aligned}
&Pr\left(\frac{\hat{\overline{Y}} - Y^*}{\sqrt{V\left(\hat{\overline{Y}}\right)}} > z_{1-\alpha} \,\middle|\, \overline{Y} = Y^* + d\right) \\
&= Pr\left(\frac{\hat{\overline{Y}} - Y^*}{\sqrt{\frac{Y^* \cdot (1 - Y^*)}{n}}} > z_{1-\alpha} \,\middle|\, \overline{Y} = Y^* + d\right) \\
&= Pr\left(\hat{\overline{Y}} - Y^* > z_{1-\alpha} \cdot \sqrt{\frac{Y^* \cdot (1 - Y^*)}{n}} \,\middle|\, \overline{Y} = Y^* + d\right)
\end{aligned}$$

What we need to do is maneuver somehow to obtain, as in the earlier derivation, a statistic that follows a standard normal distribution under the alternative hypothesis.

Once again, we will resort to a little trickery (but, as before, legitimate trickery) to do this. First, we subtract d from both sides to obtain

$$Pr\left(\hat{\overline{Y}} - Y^* > z_{1-\alpha} \cdot \left(\sqrt{\frac{Y^* \cdot (1 - Y^*)}{n}}\right) \,\middle|\, \overline{Y} = Y^* + d\right)$$

$$= Pr\left(\hat{\overline{Y}} - Y^* - d > z_{1-\alpha} \cdot \sqrt{\frac{Y^* \cdot (1 - Y^*)}{n}} - d \,\middle|\, \overline{Y} = Y^* + d\right)$$

We then divide both sides by

$$\sqrt{\frac{Y_a \cdot (1 - Y_a)}{n}}$$

(where, once again, $Y_a = Y^* + d$ for notational simplicity), leaving us with

$$Pr\left(\hat{\overline{Y}} - Y^* - d > z_{1-\alpha} \cdot \sqrt{\frac{Y^* \cdot (1 - Y^*)}{n}} - d \,\middle|\, \overline{Y} = Y^* + d\right)$$

$$= Pr\left(\frac{\hat{\overline{Y}} - Y^* - d}{\sqrt{\frac{Y_a \cdot (1-Y_a)}{n}}} > \frac{z_{1-\alpha} \cdot \sqrt{\frac{Y^* \cdot (1-Y^*)}{n}} - d}{\sqrt{\frac{Y_a \cdot (1-Y_a)}{n}}} \,\middle|\, \overline{Y} = Y^* + d\right)$$

However,

$$\frac{\hat{\overline{Y}} - Y^* - d}{\sqrt{\frac{Y_a \cdot (1-Y_a)}{n}}}$$

follows a standard normal distribution under the alternative hypothesis. Hence we have

$$Pr\left(\frac{\hat{\overline{Y}} - Y^* - d}{\sqrt{\frac{Y_a \cdot (1-Y_a)}{n}}} > \frac{z_{1-\alpha} \cdot \sqrt{\frac{Y^* \cdot (1-Y^*)}{n}} - d}{\sqrt{\frac{Y_a \cdot (1-Y_a)}{n}}} \,\middle|\, \overline{Y} = Y^* + d\right)$$

$$= Pr\left(Z > \frac{z_{1-\alpha} \cdot \sqrt{\frac{Y^* \cdot (1-Y^*)}{n}} - d}{\sqrt{\frac{Y_a \cdot (1-Y_a)}{n}}} \,\middle|\, \overline{Y} = Y^* + d\right)$$

This is the power function for the test.

We want to achieve a power of $1 - \beta$. Thus, we want

$$Pr\left(Z > \frac{z_{1-\alpha} \cdot \sqrt{\frac{Y^* \cdot (1-Y^*)}{n}} - d}{\sqrt{\frac{Y_a \cdot (1-Y_a)}{n}}} \,\middle|\, \overline{Y} = Y^* + d\right)$$

$$= Pr\left(Z > -z_{1-\beta} \,\middle|\, \overline{Y} = Y^* + d\right) = 1 - \beta$$

To find the sample size n, we thus set

$$\frac{z_{1-\alpha} \cdot \sqrt{\frac{Y^* \cdot (1-Y^*)}{n}} - d}{\sqrt{\frac{Y_a \cdot (1-Y_a)}{n}}} = -z_{1-\beta}$$

and solve for n. First we have

$$z_{1-\alpha} \cdot \sqrt{\frac{Y^* \cdot (1 - Y^*)}{n}} + z_{1-\beta} \cdot \sqrt{\frac{Y_a \cdot (1 - Y_a)}{n}} = d$$

or equivalently

$$\frac{1}{\sqrt{n}} \cdot \left[z_{1-\alpha} \cdot \sqrt{Y^* \cdot (1 - Y^*)} + z_{1-\beta} \cdot \sqrt{Y_a \cdot (1 - Y_a)} \right] = d$$

Next, we have

$$\frac{1}{\sqrt{n}} = \frac{d}{z_{1-\alpha} \cdot \sqrt{Y^* \cdot (1 - Y^*)} + z_{1-\beta} \cdot \sqrt{Y_a \cdot (1 - Y_a)}}$$

Inverting both sides and squaring we have

$$n = \left[\frac{z_{1-\alpha} \cdot \sqrt{Y^* \cdot (1 - Y^*)} + z_{1-\beta} \cdot \sqrt{Y_a \cdot (1 - Y_a)}}{d} \right]^2$$

Clearly, sample size is increasing as α and β get smaller (which, in the latter case, means as power $1 - \beta$ gets larger) and as the minimum difference d between the null and alternative that we wish to detect gets smaller. Recalling Figure 3.2 (which plotted $\overline{Y} \cdot (1 - \overline{Y})$ for a generic proportion $\overline{Y}$), it should also be clear that the sample size will be larger the closer are Y^* and Y_a to 0.5.

This sample size estimation can be performed using the STATA command `power` by specifying:

```
power oneproportion p0 pa [, alpha(numlist) power(numlist) onesided]
```

where `oneproportion` stands for a one-sample proportion test, and `p0` and `pa` denote Y^* and $Y^* + d = Y_a$, respectively. In the options, we specify the significance level (α) in `alpha(numlist)`, and the power level $(1 - \beta)$in `power(numlist)`. We specify `onesided` to perform a one-sided test.[27]

What remains is to explore the case of a two-sided test and testing for a difference in parameters between two populations with a sample from each population. We begin with the case of a two-sided test. For the two-sided test case, we will actually explicitly solve for the sample size n only for the case of continuous $\overline{Y}$ (the solution in the case where $\overline{Y}$ is a proportion is a natural application of the logic of sample size estimation for one-sided tests of proportions that we have just discussed, to the solution approach for two-sided tests that we are about to discuss). Once again, for the two-sided discussion we stick to the one-sample case.

To begin with, we must state our null and, in the most general terms at this point, the alternative hypothesis. The null hypothesis is still

$$H_0 : \overline{Y} = Y^*$$

while the alternative for a two-sided test would be

$$H_a : \overline{Y} \neq Y^*$$

[27] Alternatively, we can perform the same sample size estimation by specifying:

```
power oneproportion p0 [, diff(numlist) alpha(numlist) power(numlist) onesided]
```

where `diff(numlist)` denotes d.

Notice that this alternative is broader than the old alternative

$$H_a : \overline{Y} > Y^*$$

which drove thinking in the one-sided test case.

This newer alternative

$$H_a : \overline{Y} \neq Y^*$$

is broader because it can obtain in two ways: if $\overline{Y}$ exceeds Y^* or if $\overline{Y}$ falls short of Y^*. This means that we have to be concerned with both tails of the distribution of the test statistic t under the null hypothesis for purposes of deciding whether to accept or reject the null hypothesis.

Once again we begin with the distribution of the test statistic under the null hypothesis

$$H_0 : \overline{Y} = Y^*$$

And once again we rely on the standard normal distribution for our critical values. In the one-sided test case, the critical value was $z_{1-\alpha}$ because all of the probability of a Type-I error was focused on the possibility that we would witness

$$t > c = z_{1-\alpha}$$

even if the null hypothesis was true. In other words, we could commit a Type-I error only if the test statistic t was really big when the null hypothesis H_0 was true.

The challenge in a two-sided test with an alternative hypothesis like

$$H_a = \overline{Y} \neq Y^*$$

is that there are two ways that a Type-I error can occur. A Type-I error would occur if either

$$t < -c$$

or

$$t > c$$

when the null hypothesis H_0 is true. Because of this, the probability of a Type-I error, α, must be split between these two possibilities. The critical values for the two cases would then be

$$t < -c = -z_{1-\alpha/2}$$

or

$$t > c = z_{1-\alpha/2}$$

This basic situation of a two-sided test is illustrated in Figure 3.8.

To move to sample size estimation, we need to consider some distance d from the null hypothesis value of Y^* that we wish to detect. This is the distance on either side of Y^* that we wish to detect with power $1-\beta$. Thus, in essence, our alternative hypothesis involves two possibilities:

$$H_a : \overline{Y} \leq Y^* - d \text{ or } \overline{Y} \geq Y^* + d$$

To operationalize the derivation of a sample size estimate, we would focus on the specific alternative possibilities:

$$\overline{Y} = Y^* + d$$

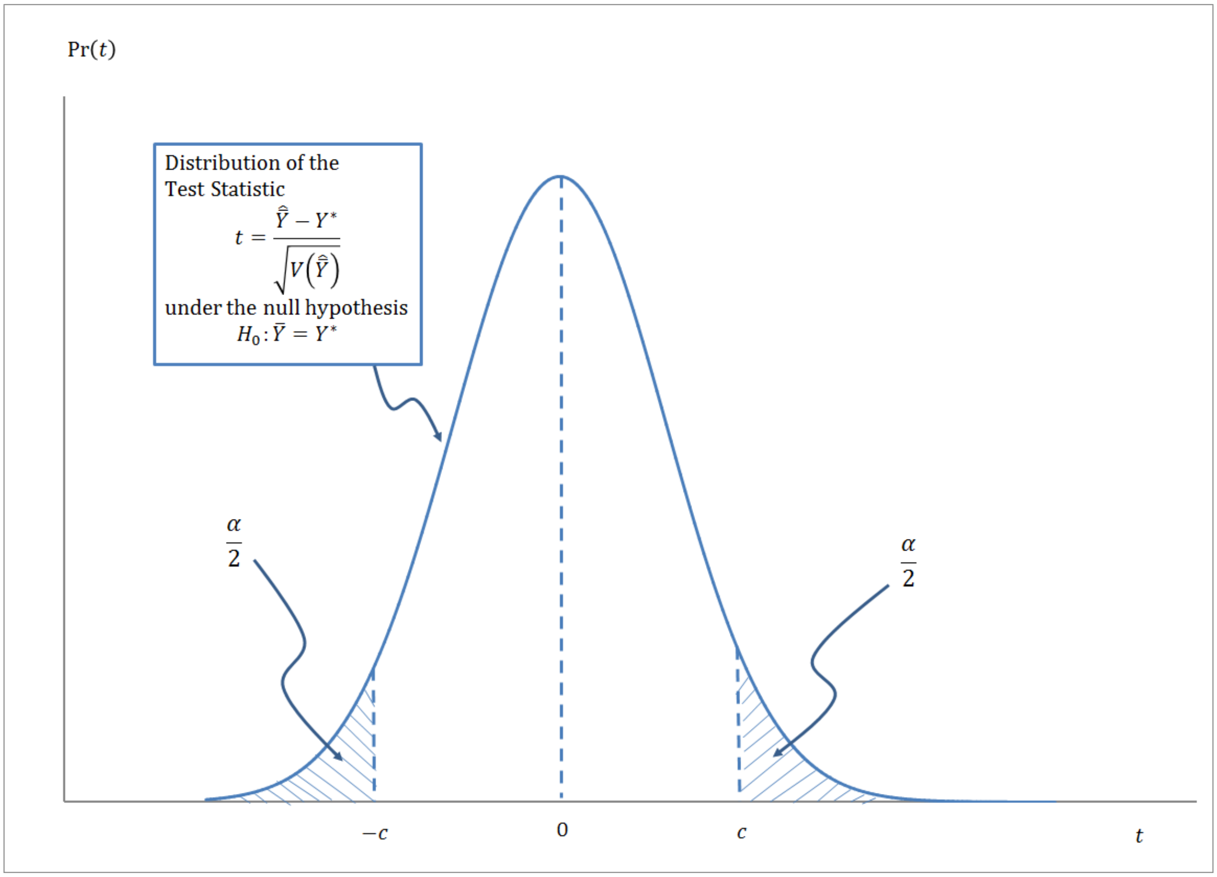
and

$$\overline{Y} = Y^* - d$$

In principle, we would need to derive sample sizes to satisfy the power requirement in the face of both of these specific alternatives. Luckily, these two roads lead to the same result.

We therefore derive for the case of the specific alternative $\overline{Y} = Y^* + d$. We begin with the probability that we would accept the null hypothesis H_0 even when $\overline{Y} = Y^* + d$ was true. This is

$$Pr\left(|t| \leq z_{1-\alpha/2} \,\middle|\, \overline{Y} = Y^* + d\right)$$



Pr(t)
Distribution of the Test Statistic
$t = \frac{\hat{\overline{Y}} - Y^*}{\sqrt{V(\hat{\overline{Y}})}}$
under the null hypothesis
$H_0 : \overline{Y} = Y^*$
$\frac{\alpha}{2}$
$\frac{\alpha}{2}$
$-c$
0
c
t

Figure 3.8. Significance in a two-sided test

Notice that the critical value $z_{1-\alpha/2}$ is crafted in a fashion that recognizes the two ways that a Type-I error can occur. This can be recast as

$$Pr\left(-z_{1-\alpha/2} \leq \frac{\hat{\overline{Y}} - Y^*}{\sqrt{V\left(\hat{\overline{Y}}\right)}} \leq z_{1-\alpha/2} \,\middle|\, \overline{Y} = Y^* + d\right)$$

Subtracting

$$\frac{d}{\sqrt{V\left(\hat{\overline{Y}}\right)}}$$

across this chain of inequalities, we have

$$Pr\left(-z_{1-\alpha/2} - \frac{d}{\sqrt{V\left(\hat{\overline{Y}}\right)}} \leq \frac{\hat{\overline{Y}} - Y^*}{\sqrt{V\left(\hat{\overline{Y}}\right)}} - \frac{d}{\sqrt{V\left(\hat{\overline{Y}}\right)}} \leq z_{1-\alpha/2} - \frac{d}{\sqrt{V\left(\hat{\overline{Y}}\right)}} \,\middle|\, \overline{Y} = Y^* + d\right)$$

Notice that the center of this chain of inequalities

$$\frac{\hat{\overline{Y}} - Y^*}{\sqrt{V\left(\hat{\overline{Y}}\right)}} - \frac{d}{\sqrt{V\left(\hat{\overline{Y}}\right)}}$$

follows the standard normal distribution under the alternative $\overline{Y} = Y^* + d$.

We thus have

$$Pr\left(-z_{1-\alpha/2} - \frac{d}{\sqrt{V\left(\hat{\overline{Y}}\right)}} \leq \frac{\hat{\overline{Y}} - Y^*}{\sqrt{V\left(\hat{\overline{Y}}\right)}} - \frac{d}{\sqrt{V\left(\hat{\overline{Y}}\right)}} \leq z_{1-\alpha/2} - \frac{d}{\sqrt{V\left(\hat{\overline{Y}}\right)}} \,\middle|\, \overline{Y} = Y^* + d\right)$$

$$= Pr\left(-z_{1-\alpha/2} - \frac{d}{\sqrt{V\left(\hat{\overline{Y}}\right)}} \leq Z \leq z_{1-\alpha/2} - \frac{d}{\sqrt{V\left(\hat{\overline{Y}}\right)}} \,\middle|\, \overline{Y} = Y^* + d\right)$$

This is simply the space under the standard normal distribution between

$$-z_{1-\alpha/2} - \frac{d}{\sqrt{V\left(\hat{\overline{Y}}\right)}}$$

and

$$z_{1-\alpha/2} - \frac{d}{\sqrt{V\left(\hat{\overline{Y}}\right)}}$$

However, that is simply the cumulative probability of z to the second point minus the cumulative probability of z to the first point. This can thus be re-written

$$Pr\left(-z_{1-\alpha/2}-\frac{d}{\sqrt{V\left(\hat{\overline{Y}}\right)}}\leq Z\leq z_{1-\alpha/2}-\frac{d}{\sqrt{V\left(\hat{\overline{Y}}\right)}}\;\middle|\;\overline{Y}=Y^{*}+d\right)$$

$$=Pr\left(Z\leq z_{1-\alpha/2}-\frac{d}{\sqrt{V\left(\hat{\overline{Y}}\right)}}\;\middle|\;\overline{Y}=Y^{*}+d\right)-Pr\left(Z\leq -z_{1-\alpha/2}-\frac{d}{\sqrt{V\left(\hat{\overline{Y}}\right)}}\;\middle|\;\overline{Y}=Y^{*}+d\right)$$

Notice, however, that the second term

$$Pr\left(Z\leq -z_{1-\alpha/2}-\frac{d}{\sqrt{V\left(\hat{\overline{Y}}\right)}}\;\middle|\;\overline{Y}=Y^{*}+d\right)$$

should roughly equal 0: there is very little probability of z having such small values if $\overline{Y}=Y^{*}+d$ (let alone a larger value for $\overline{Y}$). We can thus simplify:

$$Pr\left(Z\leq z_{1-\alpha/2}-\frac{d}{\sqrt{V\left(\hat{\overline{Y}}\right)}}\;\middle|\;\overline{Y}=Y^{*}+d\right)-Pr\left(Z\leq -z_{1-\alpha/2}-\frac{d}{\sqrt{V\left(\hat{\overline{Y}}\right)}}\;\middle|\;\overline{Y}=Y^{*}+d\right)$$

$$=Pr\left(Z\leq z_{1-\alpha/2}-\frac{d}{\sqrt{V\left(\hat{\overline{Y}}\right)}}\;\middle|\;\overline{Y}=Y^{*}+d\right)$$

So this long mathematical journey has left us with

$$Pr\left(|t|\leq z_{1-\alpha/2}\;\middle|\;\overline{Y}=Y^{*}+d\right)=Pr\left(Z\leq z_{1-\alpha/2}-\frac{d}{\sqrt{V\left(\hat{\overline{Y}}\right)}}\;\middle|\;\overline{Y}=Y^{*}+d\right)$$

We are now in a position to talk about power. Power is simply

$$Pr\left(|t|>z_{1-\alpha/2}\;\middle|\;\overline{Y}=Y^{*}+d\right)$$

But this is simply

$$Pr\left(|t|>z_{1-\alpha/2}\;\middle|\;\overline{Y}=Y^{*}+d\right)=1-Pr\left(|t|\leq z_{1-\alpha/2}\;\middle|\;\overline{Y}=Y^{*}+d\right)$$

$$= 1 - Pr\left(Z \leq z_{1-\alpha/2} - \frac{d}{\sqrt{V\left(\hat{\overline{Y}}\right)}} \,\middle|\, \overline{Y} = Y^* + d \right)$$

$$= Pr\left(Z > z_{1-\alpha/2} - \frac{d}{\sqrt{V\left(\hat{\overline{Y}}\right)}} \,\middle|\, \overline{Y} = Y^* + d \right)$$

This is the power function for the test. If we want a power of $1 - \beta$ this implies that

$$z_{1-\alpha/2} - \frac{d}{\sqrt{V\left(\hat{\overline{Y}}\right)}} = -z_{1-\beta}$$

For the continuous case we would use

$$V\left(\hat{\overline{Y}}\right) = \frac{\sigma_Y^2}{n}$$

Then we have

$$z_{1-\alpha/2} - \frac{d}{\sqrt{\frac{\sigma_Y^2}{n}}} = -z_{1-\beta}$$

or

$$z_{1-\alpha/2} + z_{1-\beta} = \frac{d}{\sqrt{\frac{\sigma_Y^2}{n}}}$$

leading to

$$\sqrt{\frac{\sigma_Y^2}{n}} = \frac{d}{z_{1-\alpha/2} + z_{1-\beta}}$$

and then

$$\frac{1}{\sqrt{n}} = \frac{d}{\sigma_Y \cdot \left(z_{1-\alpha/2} + z_{1-\beta}\right)}$$

Inverting both sides and squaring, we have

$$n = \left[\frac{\sigma_Y \cdot \left(z_{1-\alpha/2} + z_{1-\beta}\right)}{d} \right]^2$$

So basically a great deal of math has led us to essentially exactly the same place as the one-sided test case except that now the critical value is $z_{1-\alpha/2}$ rather than $z_{1-\alpha}$.

We could have premised the journey on the other alternative hypothesis

$$\overline{Y} = Y^* - d$$

being true. This would have involved

$$Pr\left(|t| \leq z_{1-\alpha/2} \,\middle|\, \overline{Y} = Y^* - d \right)$$

$$= Pr\left(-z_{1-\alpha/2} \leq \frac{\hat{\overline{Y}} - Y^*}{\sqrt{V\left(\hat{\overline{Y}}\right)}} \leq z_{1-\alpha/2} \,\middle|\, \overline{Y} = Y^* - d\right)$$

We would then *add*

$$\frac{d}{\sqrt{V\left(\hat{\overline{Y}}\right)}}$$

across the chain of inequalities. This would eventually lead to

$$Pr\left(|t| \leq z_{1-\alpha/2} \,\middle|\, \overline{Y} = Y^* - d\right)$$

$$= Pr\left(Z \leq z_{1-\alpha/2} + \frac{d}{\sqrt{V\left(\hat{\overline{Y}}\right)}} \,\middle|\, \overline{Y} = Y^* - d\right) - Pr\left(Z \leq -z_{1-\alpha/2} + \frac{d}{\sqrt{V\left(\hat{\overline{Y}}\right)}} \,\middle|\, \overline{Y} = Y^* - d\right)$$

The trick to getting rid of one of these terms this time is to recognize that

$$Pr\left(Z \leq z_{1-\alpha/2} + \frac{d}{\sqrt{V\left(\hat{\overline{Y}}\right)}} \,\middle|\, \overline{Y} = Y^* - d\right)$$

is essentially 1 if $\overline{Y} = Y^* - d$. The result of the solution proceeds analogously to the case where $\overline{Y} = Y^* + d$. We end up with

$$Pr\left(z \leq -z_{1-\alpha/2} + \frac{d}{\sqrt{V\left(\hat{\overline{Y}}\right)}} \,\middle|\, \overline{Y} = Y^* - d\right) = 1 - \beta$$

Setting

$$-z_{1-\alpha/2} + \frac{d}{\sqrt{V\left(\hat{\overline{Y}}\right)}} = z_{1-\beta}$$

we can solve for the same n.

The STATA command `power` that we used in the sample size estimation for a one-sided test can perform the sample size estimation for a two-sided test. The difference in the command specification is that we do not specify `onesided` (which we specified for a one-sided test case) for a two-sided test. This is because the default of the STATA command `power` is a two-sided test. Therefore the basic syntax of the command is:

```
   power onemean m0 ma [, alpha(numlist) power(numlist) sd(numlist)
knownsd]
```

where `onemean` stands for a one-sample mean test, and `m0` and `ma` denote $\overline{Y}$ and $\overline{Y} + d$, respectively. In the options, we specify the significance level (α) in `alpha(numlist)`, the power level

$(1-\beta)$ in power(numlist), and the standard deviation (σ_Y) in sd(numlist). The formula is based on the assumption that the standard deviation is known, so we specify knownsd.[28]

The case where $\overline{Y}$ is a proportion follows similar grueling math but similar simple logic. To solve for n we simply use the variances for $\hat{\overline{Y}}$ under the null and alternative in the one-sided cases. And, like the continuous outcome $\overline{Y}$, the sample size formula for n that obtained for proportions in the one sample case with a one-sided test emerges in the two-sided case as well, except that we substitute $z_{1-\alpha/2}$ for z_α.

The STATA command power can perform the sample size estimation when $\overline{Y}$ is a proportion and the test is two-sided. Again the difference in the command specifications from the one-sided proportion test is that we do not specify onesided (which we specified for a one-sided test case) for a two-sided test. The command for a two-sided proportion test is therefore:

```
power oneproportion p0 pa [, alpha(numlist) power(numlist)]
```

where oneproportion stands for a one-sample proportion test, and p0 and pa denote Y^* and Y^*+d, respectively. In the options, we specify the significance level (α) in alpha(numlist), and the power level $(1-\beta)$in power(numlist).[29]

Two-sided tests are expensive in terms of required sample size. Since

$$z_{1-\alpha} < z_{1-\alpha/2}$$

we have, all other things being equal,

$$n_{Two-sided} = \left[\frac{\sigma_Y \cdot \left(z_{1-\alpha/2} + z_{1-\beta}\right)}{d}\right]^2$$

$$> \left[\frac{\sigma_Y \cdot (z_{1-\alpha} + z_{1-\beta})}{d}\right]^2 = n_{One-sided}$$

Let us consider a numerical example. Suppose that we wish to perform a one-sample test of the null hypothesis $H_0 : \overline{Y} = 0.45$ with $d = 0.05$, $1-\beta = .8$ and $\alpha = 0.05$. The one-sided sample size is $n = 600$, while the two-sided sample size is $n = 761$.[30]

The sample size required for the two-sided test is thus nearly 27 percent larger than that required for the one-sided test, a difference that is nothing to sniff at in the face of tight survey resources.

In light of this, one might ask whether one- or two-sided tests should be the objective in sample

[28] Alternatively, the same sample size estimation can be performed by specifying:

```
power onemean m0[, diff(numlist) alpha(numlist) power(numlist) sd(numlist) knownsd]
```

where diff(numlist) denotes d.

[29] Alternatively, we can perform the same sample size estimation by specifying:

```
power oneproportion p0 [, diff(numlist) alpha(numlist) power(numlist)]
```

where diff(numlist) denotes d.

[30] To calculate this we used the STATA code:

```
power oneproportion 0.4 0.45, alpha(0.05) power(0.8) onesided
```

for the one-sided test and:

```
power oneproportion 0.4 0.45, alpha(0.05) power(0.8)
```

for the two-sided test.

size estimation for program impact evaluations. We would argue forcefully that two-sided tests should be the objective, and for at least two reasons. First, the null hypothesis in program impact evaluation is typically "no impact." Since the critical value in a one-sided test is smaller, the analyst is in some sense putting their fingers on the scale in terms of making it easier to show impact. That would be fine if the consumers of program impact evaluation results, those who need to make decisions about programmatic investments, understood that fully. Too often they don't realize that showing impact in a one-sided test is not like showing it in a two-sided test: in the latter, the burden of statistical evidence is much higher. One can alternatively set lower levels for α in one-sided tests.

Second, and perhaps more importantly, programs can actually prove counterproductive. The findings from program impact evaluations tell us something about what works, and what doesn't, so that we can march toward a better tomorrow where scarce resources are allocated more efficiently and effectively so that even higher levels of human welfare can be achieved with available resources. That is the ultimate, global social purpose of program impact evaluations. We need to understand fully the implications of programs, including whether they actually did harm. We should not rely on testing that by construction allows us to ignore really unpleasant potential realities.

Finally, we turn to the case where we compare estimates from two different populations. This requires a bit of (temporary) expansion of our notation. We indicate the population with a subscript in the indicator:

$$\overline{Y}_j$$

where j indicates population. For this discussion, we will consider two populations, Y_1 and Y_2. This approach to sampling is motivated by the idea that we will be selecting samples from two different populations, estimating an indicator Y for each of them, and then performing tests for differences in the value of $\overline{Y}$ between the populations through those estimates. The estimates from the two populations are $\hat{\overline{Y}}_1$ and $\hat{\overline{Y}}_2$. To begin with, we have the null hypothesis

$$H_0 : \overline{Y}_1 - \overline{Y}_2 = Y^*$$

This would give rise to the test statistic

$$t = \frac{\hat{\overline{Y}}_1 - \hat{\overline{Y}}_2 - Y^*}{\sqrt{V\left(\hat{\overline{Y}}_1 - \hat{\overline{Y}}_2\right)}}$$

We accept[31] the null hypothesis at the α level of significance in the one-sided test case if

$$t \leq z_{1-\alpha}$$

[31] Again, people often say something like "fail to reject" instead of accept because few would be prepared to assert that $\overline{Y}_1 - \overline{Y}_2$ exactly equals Y^*. We often say "accept" because it is too cumbersome to keep using "fail to reject."

while the acceptance rule for the two-sided test is

$$|t| \leq z_{1-\alpha/2}$$

This is the basic design of the benchmark framework for testing for differences in indicators.

As such it has clear and immediate potential application to impact evaluation. Specifically, it is natural to think of population 1 as program participants and population 2 as a population of non-participant comparison units. Impact would then be

$$\overline{Y}_1 - \overline{Y}_2$$

We often want to test whether a program had some kind of influence against a null that it had none. This would imply setting Y^* to zero. As we work through this case we won't do that but instead leave the explicit term Y^* in the derivations. There are two reasons we do this. First, we do so simply so that the reader can see how a $Y^* > 0$ would wend its way through the derivation process. Second, even in the evaluation context it is possible one would want to assume a $Y^* > 0$. For instance, if one decided to rely on an explicitly pre-/post-design, $\overline{Y}_1$ might represent $\overline{Y}$ at the follow-up "post" observation point, and $\overline{Y}_2$ the same for the initial "pre" observation point. A $Y^* > 0$ then might capture some secular trend to outcomes that the researcher feels would have happened without the program, in which case $\overline{Y}_1 - \overline{Y}_2 = Y^*$ would represent no program impact beyond what the trend would have done anyway.

The general form alternative hypothesis in this case is

$$H_a : \overline{Y}_1 - \overline{Y}_2 \neq Y^*$$

To actually operationalize sample size estimation we would again need to get a bit more specific about how much difference between $\overline{Y}_1 - \overline{Y}_2$ and H^* we wished to detect with $1 - \beta$ power. To do that we would need to specify how big a deviation from H^* was important to us for the purpose of establishing power. As before, we specify that difference as d.

The alternative for a one-sided test would then be

$$H_a : \overline{Y}_1 - \overline{Y}_2 \geq Y^* + d$$

while for a two-sided test we would have

$$H_a : \overline{Y}_1 - \overline{Y}_2 \leq Y^* - d \text{ or } \overline{Y}_1 - \overline{Y}_2 \geq Y^* + d$$

As we have already mentioned, the main difference between the one- and two-sided cases is again how the critical value is handled: in the one-sided case it would be $z_{1-\alpha}$ while for two-sided tests it would be $z_{1-\alpha/2}$.

The derivation of a power function and sample size estimate must overcome one issue: characterizing the variance

$$V\left(\hat{\overline{Y}}_1 - \hat{\overline{Y}}_2\right)$$

which by its very nature will be a bit trickier than

$$V\left(\hat{\overline{Y}}\right)$$

Since the estimates $\hat{\overline{Y}}_1$ to $\hat{\overline{Y}}_2$ are (*ex-ante*) random variables,

$$V\left(\hat{\overline{Y}}_1 - \hat{\overline{Y}}_2\right) = V\left(\hat{\overline{Y}}_1\right) + V\left(\hat{\overline{Y}}_1\right) - 2 \cdot cov\left(\hat{\overline{Y}}_1, \hat{\overline{Y}}_2\right)$$

where

$$cov\left(\hat{\overline{Y}}_1, \hat{\overline{Y}}_2\right) = \frac{1}{n} \cdot \rho \cdot \sigma_{Y_1}\sigma_{Y_2}$$

for

$$\rho = corr\left(\overline{Y}_1, \overline{Y}_2\right)$$

that is, the correlation between $\overline{Y}_1$ and $\overline{Y}_2$.

One way that

$$cov\left(\hat{\overline{Y}}_1, \hat{\overline{Y}}_2\right) \neq 0$$

is if there is some kind of fundamental correlation in the observations. Consider, for instance, a pre-/post-setup where there can be shocks that affect $\overline{Y}$ over multiple time periods. Another way that this can emerge is if there is longitudinal data or if one has exclusively cross-sectional data but the populations are not completely unrelated to one another (as in the case where, for instance, a multistage design participants and non-participants come from the same clusters).

One neat trick[32] is to assume that

$$\sigma_{Y_1}^2 = \sigma_{Y_2}^2 = \sigma_Y^2$$

Then

$$\begin{aligned} V\left(\hat{\overline{Y}}_1 - \hat{\overline{Y}}_2\right) &= V\left(\hat{\overline{Y}}_1\right) + V\left(\hat{\overline{Y}}_2\right) - 2 \cdot cov\left(\hat{\overline{Y}}_1, \hat{\overline{Y}}_2\right) \\ &= \frac{\sigma_{Y_1}^2}{n} + \frac{\sigma_{Y_2}^2}{n} - 2 \cdot cov\left(\hat{\overline{Y}}_1, \hat{\overline{Y}}_2\right) \\ &= \frac{\sigma_Y^2}{n} + \frac{\sigma_Y^2}{n} - 2 \cdot cov\left(\hat{\overline{Y}}_1, \hat{\overline{Y}}_2\right) \\ &= 2 \cdot \frac{\sigma_Y^2}{n} - 2 \cdot cov\left(\hat{\overline{Y}}_1, \hat{\overline{Y}}_2\right) \end{aligned}$$

[32] Neat is good; legitimate is even better. This trick is one that Deaton (2010) would probably scoff at in the context of randomized controlled trials (more on this in the next section). That said, in practice it might not be so problematic: we would be employing it from the *ex-ante* perspective, from which we would be lucky to have any prior for the variance, let alone separate priors for the two different populations.

$$= 2 \cdot \frac{\sigma_Y^2}{n} - 2 \cdot \frac{\sigma_Y^2}{n} \cdot \frac{cov\left(\hat{\overline{Y}}_1, \hat{\overline{Y}}_2\right)}{\frac{\sigma_Y^2}{n}}$$

Next, we have the result that

$$cov\left(\hat{\overline{Y}}_1, \hat{\overline{Y}}_2\right) = \frac{1}{n} \cdot \rho \cdot \sigma_{Y_1}\sigma_{Y_2}$$

$$= \frac{1}{n} \cdot \rho \cdot \sigma_Y^2$$

by our departing assumption. Then, continuing, we have

$$V\left(\hat{\overline{Y}}_1 - \hat{\overline{Y}}_2\right) = 2 \cdot \frac{\sigma_Y^2}{n} - 2 \cdot \frac{\sigma_Y^2}{n} \cdot \frac{\frac{1}{n} \cdot \rho \cdot \sigma_Y^2}{\frac{\sigma_Y^2}{n}}$$

$$= 2 \cdot \frac{\sigma_Y^2}{n} - 2 \cdot \frac{\sigma_Y^2}{n} \cdot \rho$$

Thus

$$V\left(\hat{\overline{Y}}_1 - \hat{\overline{Y}}_2\right) = 2 \cdot \frac{\sigma_Y^2}{n}\left(1 - \rho\right)$$

This can (as we will see) be a very handy device for capturing possible correlation of observations between the two populations.

Let us derive a power function and sample size estimate in the two-sided case. We begin with the test statistic

$$t = \frac{\hat{\overline{Y}}_1 - \hat{\overline{Y}}_2 - Y^*}{\sqrt{V\left(\hat{\overline{Y}}_1 - \hat{\overline{Y}}_2\right)}}$$

the notation of which we simplify by writing $\Delta\hat{\overline{Y}} = \hat{\overline{Y}}_1 - \hat{\overline{Y}}_2$. The term Δ usually indicates "difference" or "change." By similar logic, we would have $\Delta\overline{Y} = \overline{Y}_1 - \overline{Y}_2$. The test statistic would then be

$$t = \frac{\hat{\overline{Y}}_1 - \hat{\overline{Y}}_2 - Y^*}{\sqrt{V\left(\hat{\overline{Y}}_1 - \hat{\overline{Y}}_2\right)}} = \frac{\Delta\hat{\overline{Y}} - Y^*}{\sqrt{V\left(\Delta\hat{\overline{Y}}\right)}}$$

We begin with the probability of accepting the null if one side of the two-sided alternative is true:

$$Pr\left(|t| \leq z_{1-\alpha/2} \;\middle|\; \Delta\overline{Y} = Y^* + d\right)$$

$$= Pr\left(-z_{1-\alpha/2} \leq \frac{\Delta\hat{\overline{Y}} - Y^*}{\sqrt{V\left(\Delta\hat{\overline{Y}}\right)}} \leq z_{1-\alpha/2} \;\middle|\; \Delta\overline{Y} = Y^* + d\right)$$

Notice that this is exactly the same setup structurally as in the previous discussion of two-sided tests. And it is going to lead to the same place. The power function is

$$Pr\left(|t| > z_{1-\alpha/2} \;\middle|\; \Delta\overline{Y} = Y^* + d\right) = 1 - Pr\left(|t| \leq z_{1-\alpha/2} \;\middle|\; \Delta\overline{Y} = Y^* + d\right)$$

$$= 1 - Pr\left(Z \leq z_{1-\alpha/2} - \frac{d}{\sqrt{V\left(\Delta\hat{\overline{Y}}\right)}} \middle| \Delta\overline{Y} = Y^* + d \right)$$

$$= Pr\left(Z > z_{1-\alpha/2} - \frac{d}{\sqrt{V\left(\Delta\hat{\overline{Y}}\right)}} \middle| \Delta\overline{Y} = Y^* + d \right)$$

If we want the sample size, we again set

$$z_{1-\alpha/2} - \frac{d}{\sqrt{V\left(\Delta\hat{\overline{Y}}\right)}} = -z_{1-\beta}$$

and solve for n.

We do so in the continuous case. For that case we would (invoking our little trick about the variance of the difference in estimates derived above) use

$$V\left(\Delta\hat{\overline{Y}}\right) = 2 \cdot \frac{\sigma_Y^2}{n} (1 - \rho)$$

Then we have

$$z_{1-\alpha/2} - \frac{d}{\sqrt{2 \cdot \frac{\sigma_Y^2}{n} \cdot (1-\rho)}} = -z_{1-\beta}$$

or

$$z_{1-\alpha/2} + z_{1-\beta} = \frac{d}{\sqrt{2 \cdot \frac{\sigma_Y^2}{n} \cdot (1-\rho)}}$$

leading to

$$\sqrt{\frac{2 \cdot \sigma^2 \cdot (1-\rho)}{n}} = \frac{d}{z_{1-\alpha/2} + z_{1-\beta}}$$

and then

$$\frac{1}{\sqrt{n}} = \frac{d}{\sqrt{2 \cdot \sigma^2 \cdot (1-\rho)} \cdot \left(z_{1-\alpha/2} + z_{1-\beta}\right)}$$

Inverting both sides and squaring, we have

$$n = 2 \cdot (1-\rho) \left(\left[\frac{\sigma \cdot \left(z_{1-\alpha/2} + z_{1-\beta}\right)}{d} \right]^2 \right)$$

This is the sample size for *each* population. If the observations from populations 1 and 2 are uncorrelated, simply set $\rho = 0$.

The STATA command `power` can perform the sample size estimation for a test of comparing two means when $\rho = 0$ and $\sigma_{Y_1} = \sigma_{Y_2} = \sigma_Y$. The basic command specification is:

```
power twomeans m0 ma [, alpha(numlist) power(numlist) sd(numlist)
knownsd]
```

where `twomeans` stands for a two-sample mean test, and `m0` and `ma` denote $\overline{Y}_1$ and $\overline{Y}_2$, respectively. In the options, we specify the significance level (α) in `alpha(numlist)`, the power level ($1-\beta$) in `power(numlist)`, and the standard deviation (σ_Y) in `sd(numlist)`. The formula is based on the assumption that the standard deviation is known, so we specify `knownsd`. Notice that the default of the STATA command `power` is a two-sided test and hence we do not have to specify it.[33] When $\rho \neq 0$, simply multiply the estimated sample size by $(1-\rho)$ to obtain the estimated sample size accounting for ρ.

There are a couple of interesting things to note about the sample size estimate

$$n = 2 \cdot (1-\rho) \cdot \left(\left[\frac{\sigma \cdot \left(z_{1-\alpha/2} + z_{1-\beta} \right)}{d} \right]^2 \right)$$

The first is that considering testing for a difference between two populations based on estimates from those two populations itself doubles sample size per population compared with the sample size required for the one-estimate/one-population case considered in the original two-sided test discussion. This reflects the increased variance involved in testing a difference in estimates. To put this in concrete terms, suppose that one wished to perform a two-sided, one-sample test whether a population parameter equaled 55 with a d of 5 and a σ_Y of 20. The one-sample sample size would be $n = 126$. The two-sample test of whether the two population parameters differed against a possible difference d of 5 or more with a σ_Y of 20 require a sample size of 252 *from each population.*[34]

Second, having correlation between the units in the two populations can lower sample sizes, possibly by a lot. This is one reason longitudinal data are so popular: the correlations over time of outcomes work to lower necessary sample sizes. While we wholeheartedly agree with the general arguments for longitudinal data, we do have one concern from a sample size estimation perspective. Too often, particularly in the lower- and middle-income societies we work in (and, surprisingly often, in data-rich, wealthy, post-industrial societies as well) the longitudinal data infrastructure is too thin to form a truly confidence-inspiring prior assumption about ρ. Unfortunately, this information void often is filled by pressure to make extraordinarily generous assumptions about ρ, driving down estimates of required sample sizes and hence survey costs. This is too often wishful thinking by which survey designers ultimately cheat themselves.[35]

[33] Alternatively, the same sample size estimation can be performed by specifying:

```
power twomeans m0[, diff(numlist) alpha(numlist) power(numlist) sd(numlist) knownsd]
```

where `diff(numlist)` denotes $\overline{Y}_2 - \overline{Y}_1$.

[34] To calculate this, we used the STATA code

```
power onemean 55 50, alpha(0.05) power(0.8) sd(20) knownsd
```

for the one-sample test and

```
power twomeans 55 50, alpha(0.05) power(0.8) sd(20) knownsd
```

for the two-sample test.

[35] When we were graduate students, still learning "real world" sampling, we were often told that ρ values of 0.7 or 0.8 were quite reasonable. When we pressed for strong evidence for this, more often than not none was forthcoming. In general, ρ would probably depend on the indicator and, when we have been able to test, we have seen examples where the correlations were quite modest. Applied sampling is a field of endeavor where certain "stylized beliefs"

Before moving on, we briefly mention the sample size estimation formula for comparison of means in the continuous case when we forego the assumption

$$\sigma^2_{Y_1} = \sigma^2_{Y_2} = \sigma^2_Y$$

In other words, what happens when

$$\sigma^2_{Y_1} \neq \sigma^2_{Y_2}$$

We can think of two cases in this instance. First, if the estimates are uncorrelated, we have then

$$V\left(\hat{\bar{Y}}_1 - \hat{\bar{Y}}_2\right) = V\left(\Delta\hat{\bar{Y}}\right) = V\left(\hat{\bar{Y}}_1\right) + V\left(\hat{\bar{Y}}_2\right) - 2 \cdot cov\left(\hat{\bar{Y}}_1, \hat{\bar{Y}}_2\right)$$

$$= \frac{\sigma^2_{Y_1}}{n} + \frac{\sigma^2_{Y_2}}{n} - 2 \cdot 0 = \frac{\sigma^2_{Y_1}}{n} + \frac{\sigma^2_{Y_2}}{n}$$

which leads to

$$z_{1-\alpha/2} - \frac{d}{\sqrt{\frac{\left(\sigma^2_{Y_1}+\sigma^2_{Y_2}\right)}{n}}} = -z_{1-\beta}$$

or

$$z_{1-\alpha/2} + z_{1-\beta} = \frac{d}{\sqrt{\frac{\left(\sigma^2_{Y_1}+\sigma^2_{Y_2}\right)}{n}}}$$

Using the solution logic we have invoked, we derive the formula

$$n = \left(\sigma^2_{Y_1} + \sigma^2_{Y_2}\right) \cdot \left(\left[\frac{\left(z_{1-\alpha/2} + z_{1-\beta}\right)}{d}\right]^2\right)$$

The case where the two estimates are correlated is a bit trickier.

$$V\left(\hat{\bar{Y}}_1 - \hat{\bar{Y}}_2\right) = V\left(\hat{\bar{Y}}_1\right) + V\left(\hat{\bar{Y}}_2\right) - 2 \cdot cov\left(\hat{\bar{Y}}_1, \hat{\bar{Y}}_2\right)$$

$$= \frac{\sigma^2_{Y_1}}{n} + \frac{\sigma^2_{Y_2}}{n} - 2\frac{1}{n} \cdot \rho \cdot \sigma_{Y_1} \cdot \sigma_{Y_2}$$

If $\sigma^2_{Y_1}$ and $\sigma^2_{Y_2}$ aren't too far apart in value, then one can make the rough approximation

$$2 \cdot \sigma_{Y_1} \cdot \sigma_{Y_2} \approx \sigma^2_{Y_1} + \sigma^2_{Y_2}$$

can have an amazing persistence, despite mixed or even contrary evidence. It can almost bring to mind the way that many in Christopher Columbus' time clung to the belief that the world was flat, despite evidence to the contrary stretching back at least to Eratosthenes of Cyrene a millennium and a half earlier.

Then, we would have

$$V\left(\hat{\overline{Y}}_1 - \hat{\overline{Y}}_2\right) = V\left(\Delta\hat{\overline{Y}}\right) = \left(\frac{\sigma^2_{Y_1}}{n} + \frac{\sigma^2_{Y_2}}{n}\right) \cdot (1-\rho)$$

leading to

$$z_{1-\alpha/2} - \frac{d}{\sqrt{\frac{\left(\sigma^2_{Y_1}+\sigma^2_{Y_2}\right)}{n} \cdot (1-\rho)}} = -z_{1-\beta}$$

or

$$z_{1-\alpha/2} + z_{1-\beta} = \frac{d}{\sqrt{\frac{\left(\sigma^2_{Y_1}+\sigma^2_{Y_2}\right)}{n} \cdot (1-\rho)}}$$

Then the sample size estimation formula is derived

$$n = (1-\rho) \cdot \left(\sigma^2_{Y_1} + \sigma^2_{Y_2}\right) \cdot \left(\left[\frac{\left(z_{1-\alpha/2} + z_{1-\beta}\right)}{d}\right]^2\right)$$

Although this is an approximation, one must remember that approximations are implicitly or explicitly built into most sample size estimation formulas.

The STATA command `power` can perform the sample size estimation for a test of comparing two means when $\rho = 0$ and $\sigma_{Y_1} \neq \sigma_{Y_2}$. The basic command specification is:

```
power twomeans m0 ma [, alpha(numlist) power(numlist) sd1(numlist)
sd2(numlist) knownsd]
```

where `twomeans` stands for a two-sample mean test, and `m0` and `ma` denote $\overline{Y}_1$ and $\overline{Y}_2$, respectively. In the options, we specify the significance level (α) in `alpha(numlist)`, the power level ($1-\beta$) in `power(numlist)`. The difference in the command specifications from the case in which $\sigma_{Y_1} = \sigma_{Y_2}$ is that the standard deviations are specified separately, namely σ_1 and σ_2 in `sd1(numlist)` and `sd2(numlist)`, respectively. Again, the formula is based on the assumption that the standard deviation is known, so we specify `knownsd`. Notice that the default of the STATA command `power` is a two-sided test and hence we do not have to specify it.[36] When $\rho \neq 0$, simply multiply the estimated sample size by $(1-\rho)$ to obtain the estimated sample size accounting for ρ.

We conclude with a brief discussion of two-sample tests involving proportions. To simplify, suppose that the null hypothesis is

$$H_0 : \overline{Y}_1 - \overline{Y}_2 = 0$$

and the difference we wish to detect with power $1-\beta$ is d. The benchmark variance for $\hat{\overline{Y}}_1 - \hat{\overline{Y}}_2$

[36] Alternatively, the same sample size estimation can be performed by specifying:

```
power twomeans m0[, diff(numlist) alpha(numlist) power(numlist) sd1(numlist)
sd2(numlist) knownsd]
```

where `diff(numlist)` denotes $\overline{Y}_2 - \overline{Y}_1$.

under the null hypothesis is

$$\overline{\overline{Y}} \cdot \left(1 - \overline{\overline{Y}}\right) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

where $\overline{\overline{Y}}$ is the weighted average of Y defined by

$$\overline{\overline{Y}} = \frac{n_1 \cdot \overline{Y}_1 + n_2 \cdot \overline{Y}_2}{n_1 + n_2}$$

and n_1 and n_2 are the samples from populations 1 and 2, respectively. This is referred to as a "pooled estimate" and is approximately normal in large samples (Valliant, et al., 2013). If the sample sizes in the two groups are equal, this variance reduces to

$$2 \cdot \frac{\overline{\overline{Y}} \cdot \left(1 - \overline{\overline{Y}}\right)}{n}$$

where $\overline{\overline{Y}}$ now equals

$$\frac{\overline{Y}_1 + \overline{Y}_2}{2}$$

By contrast, the variance under the alternative $\overline{Y}_1 - \overline{Y}_2 = d$ is

$$\frac{\overline{Y}_1 \cdot \left(1 - \overline{Y}_1\right) + \overline{Y}_2 \left(1 - \overline{Y}_2\right)}{n}$$

With these variances in hand we have enough to derive a sample size estimate for proportions with the basic logic for doing so introduced earlier.

The end result would be

$$n = \left[\frac{z_{1-\alpha/2} \cdot \sqrt{2 \cdot \overline{\overline{Y}} \cdot \left(1 - \overline{\overline{Y}}\right)} + z_{1-\beta} \cdot \sqrt{\overline{Y}_1 \cdot \left(1 - \overline{Y}_1\right) + \overline{Y}_2 \cdot \left(1 - \overline{Y}_2\right)}}{d}\right]^2$$

One interesting question is how to handle correlation of the observations from the two populations in the case of comparing estimates of proportions. A common approach in practice is simply to multiply the sample size estimate by $(1 - \rho)$. A more nuanced discussion of correlation in the case of proportions estimates with two samples can be found in, for instance, Valliant, et al. (2013).

The STATA command `power` can perform the sample size estimation for a test of two-proportion comparison when $\rho = 0$. The basic command specification is:

```
power twoproportions p1 p2 [, alpha(numlist) power(numlist)]
```

where `twoproportions` stands for a two-sample proportion test, and `p1` and `p2` denote $\overline{Y}_1$ and $\overline{Y}_2$, respectively. In the options, we specify the significance level (α) in `alpha(numlist)`, the power level $(1 - \beta)$ in `power(numlist)`. Again, notice that the default of the STATA command `power` is a two-sided test.[37] When $\rho \neq 0$, simply multiply the estimated sample size by $(1 - \rho)$ to

[37] Alternatively, the same sample size estimation can be performed by specifying:

obtain the estimated sample size accounting for ρ.

So far, we have assumed that the samples selected from the two populations would be the same size. However, sample size estimation along the lines we have discussed can be relaxed to allow for different sample sizes from the different populations. The math of this is probably a bridge too far at this point, and so we simply present final formulas. First, for a continuous outcome, we have (from Rosner, 2006)

$$n_1 = \left[\frac{\left(\sigma^2_{\hat{\overline{Y}}_1} + \sigma^2_{\hat{\overline{Y}}_1}/r\right) \cdot \left(z_{1-\alpha/2} + z_{1-\beta}\right)}{d} \right]^2$$

and $n_2 = r \cdot n_1$. This is the sample size formula for a two-sided test. To consider a one-sided test just replace $z_{1-\alpha/2}$ with $z_{1-\alpha}$.

For the case of proportions, we turn to Fleiss, et al. (2003) who offer the following formula:

$$n_0 = \frac{\left[z_{1-\alpha/2} \cdot \sqrt{(r+1)\,\overline{\overline{Y}} \cdot \left(1 - \overline{\overline{Y}}\right)} + z_{1-\beta} \cdot \sqrt{r \cdot \overline{Y}_1 \cdot \left(1 - \overline{Y}_1\right) + \overline{Y}_2 \cdot \left(1 - \overline{Y}_2\right)}\right]^2}{r \cdot d^2}$$

where $n_1 = n_0$ and $n_2 = r \cdot n_1$. An additional layer often applied in two-sample tests of proportions involves something called a "continuity correction." One could spill a lot of ink about continuity corrections, but basically they are a factor that adjusts for our use of a sample size formula motivated by an assumption of continuity (through its reliance on the standard normal distribution) to estimate sample size for a variable that is, in fact, not continuous. The correction typically involves a step between n_0 and n_1, specifically

$$n_1 = \frac{n_0}{4} \cdot \left[1 + \sqrt{1 + \frac{2 \cdot (r+1)}{n_0 \cdot r \cdot |d|}}\right]^2$$

The continuity correction generally works to lower sample sizes a bit.

The sample size estimation for a test of two-proportion comparison when $\rho = 0$ can be performed using the STATA command `power` by specifying:

```
    power twoproportions p1 p2 [, alpha(numlist) power(numlist)
nratio(numlist) continuity]
```

which is essentially the same as the previous example with two differences: `nratio(numlist)`, in which we specify r, and `continuity`, which indicates that the continuity correction is applied. Again, when $\rho \neq 0$, simply multiply the estimated sample size by $(1 - \rho)$ to obtain the estimated sample size accounting for ρ.

Before moving on, we offer a few graphs to give some sense of the performance of the sample size estimators we have been discussing. There are so many scenarios that we could look at that we focus on just a few to give some sense of the general patterns we are confronted with when thinking

```
  power twoproportions p1 [, diff(numlist) alpha(numlist) power(numlist)]
```

where `diff(numlist)` denotes $\overline{Y}_2 - \overline{Y}_1$.

about the sample size estimates produced by these formulas. In all cases we focus on comparison of proportions, but in all cases the lessons carry over to the continuous case.

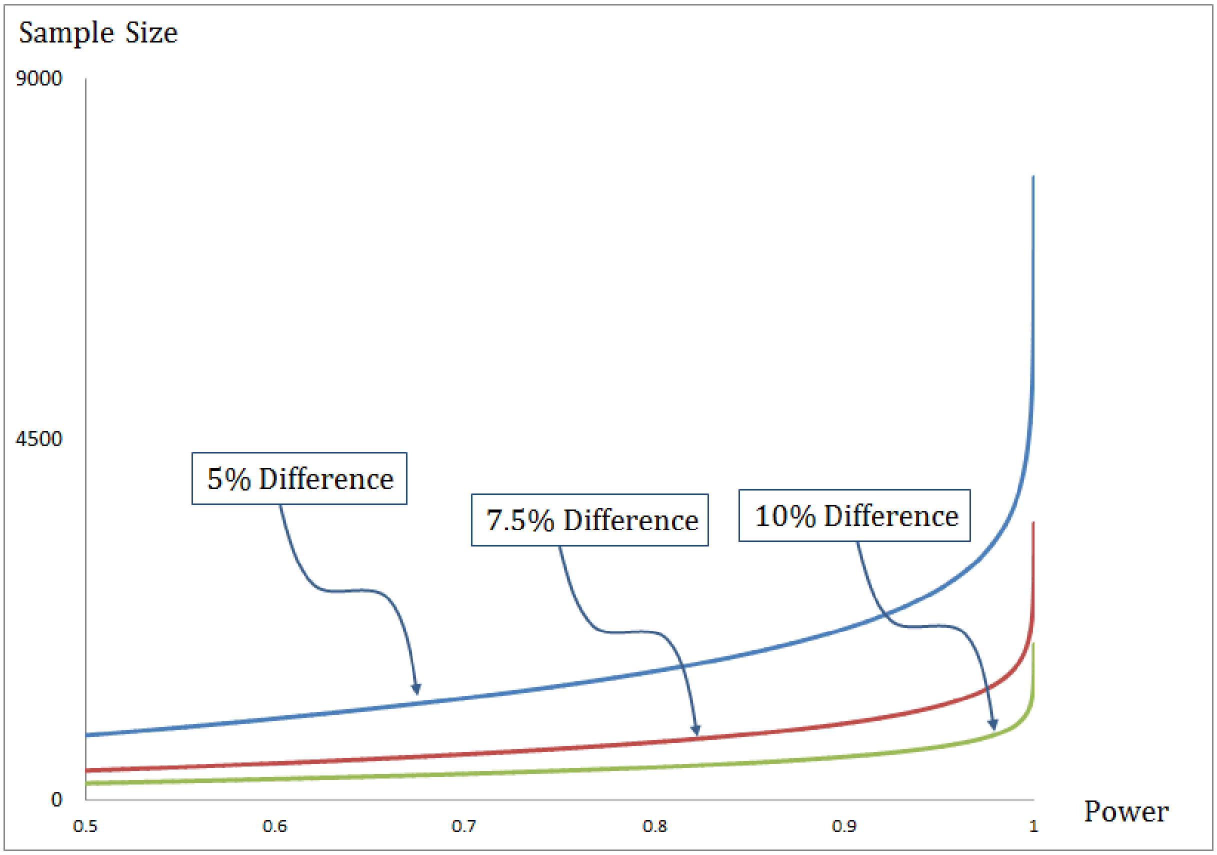


Figure 3.9. Sample sizes for various differences and power levels

In Figure 3.9 we consider how sample size varies with power. The backdrop is two-sample, two-sided tests with 5 percent significance for three scenarios:

1. $H_0 : \overline{Y}_1 = \overline{Y}_2,\ H_a : \overline{Y}_1 = 0.4750; \overline{Y}_2 : 0.5250$

2. $H_0 : \overline{Y}_1 = \overline{Y}_2,\ H_a : \overline{Y}_1 = 0.4625; \overline{Y}_2 : 0.5375$

3. $H_0 : \overline{Y}_1 = \overline{Y}_2,\ H_a : \overline{Y}_1 = 0.45; \overline{Y}_2 : 0.55$

As one can see from Figure 3.9, sample size grows nonlinearly with power, with a brisk increase becoming nearly asymptotic at the highest power values.

In Figure 3.10 we repeat the exercise for significance, considering consider how sample size varies with significance. The details of the simulation are the same as in the simulation for Figure 3.10, except that this time it is power that we hold fixed (at 80 percent) and vary significance. As significance falls, sample size increases nonlinearly. Eventually, as significance gets lower and lower, the increases in sample size required for further reductions in the significance level become increasingly massive. We thus see a pattern very similar to what we witnessed with power.

Results such as these are one reason why the authors argue against describing sample size estimation as an exercise in "powering" (as in "we are powering our sample on this indicator"). Significance has a huge influence on sample size. It seems that there is a sort of tendency to default to a significance target of 0.05, while power receives much more attention in sample size discussions (though, in the end, the power target of 80 percent nearly always prevails in the authors' experience). This is driven by two fundamentally misguided beliefs. The first is that power is the main issue of substantive importance from a scientific or behavioral standpoint (i.e., what we want to learn). Anyone who has been following the discussion in this subsection recognizes that that is absolute nonsense. Significance is incredibly important because it essentially tells us how much evidence we need to reject the null. Second, there seems to be a persistent informal myth that power is the main driver of sample size. To be sure, the choice of power level is indeed an incredibly important determinant of sample size. However, significance is also extremely important.

In some sense this is a plea for rethinking lazy conventions into which much of empirical practice has fallen. Serious thought needs to be put into the acceptable probability of rejecting a true null hypothesis or a true alternative hypothesis. Both are of great substantive interest in terms of the goals of the survey, and both have huge implications for sample size and survey resources.

Finally, in Figure 3.11 we consider the scope for having different sample sizes from the two populations. The simulation behind this figure involved a two-sample, two-sided test of proportions

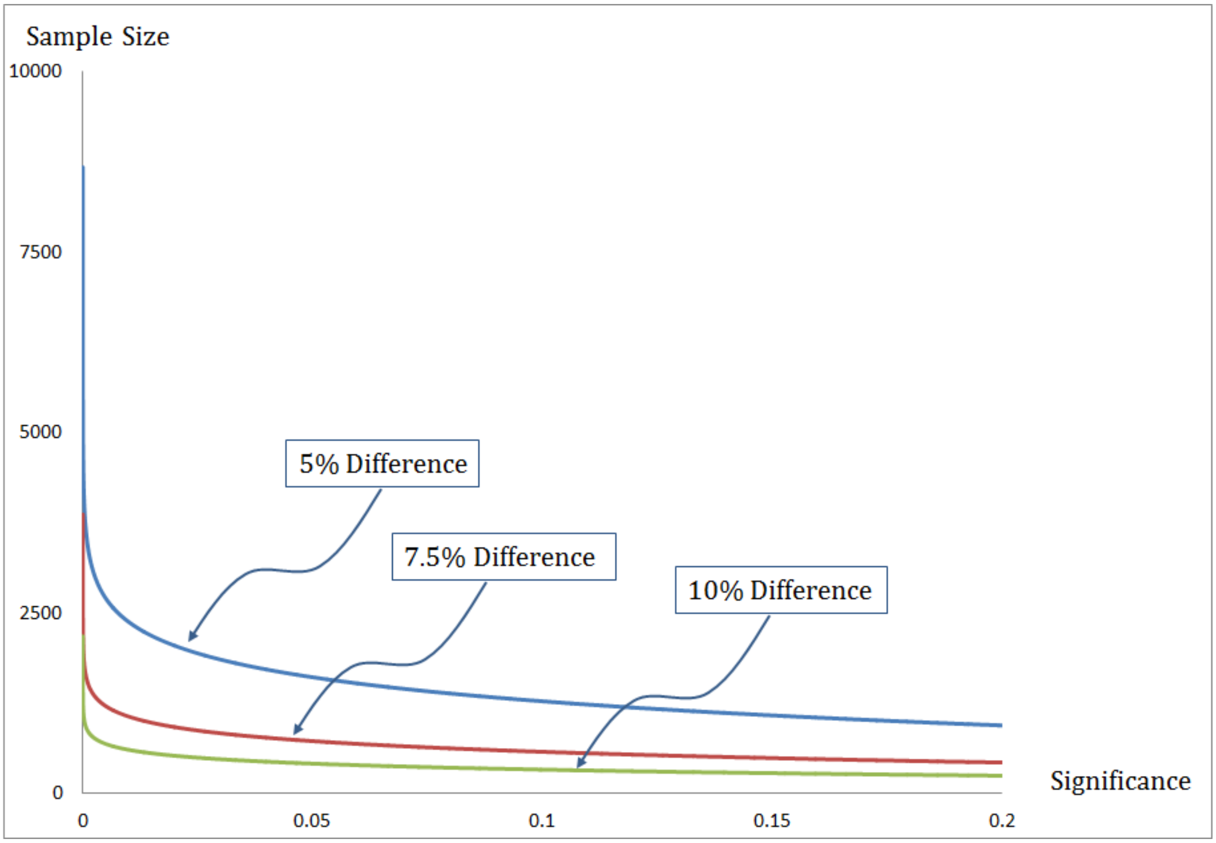


Figure 3.10. Sample sizes for various differences and significance levels

at 80 percent power and 5 percent significance of

$$H_0 : \overline{Y}_1 = \overline{Y}_2, H_a : \overline{Y}_1 = 0.4750; \overline{Y}_2 : 0.5250;$$

The variation in Figure 3.11 is that we considered various values for r where

$$n_2 = r \cdot n_1$$

The figure clearly illustrates how reductions in the size of the sample from population 1 can be offset with an increase in the size of the sample from population 2. This might be quite an attractive prospect if, for instance, conducting survey operations in population 1 was much more expensive than doing so in population 2. Another circumstance where this might be considered is where populations 1 and 2 represent repeated cross sections, and the sampling parameters for the baseline did not prove very accurate in light of baseline findings in a fashion that suggested that the originally planned baseline and follow-up sample sizes were inadequate: in that instance, one might want to compensate by "topping off" the endline sample size.

However, there is clearly a limit to the scope for this. The tradeoff becomes increasingly poor as r increases — at high values of r there are virtually no marginal reductions in the sample size

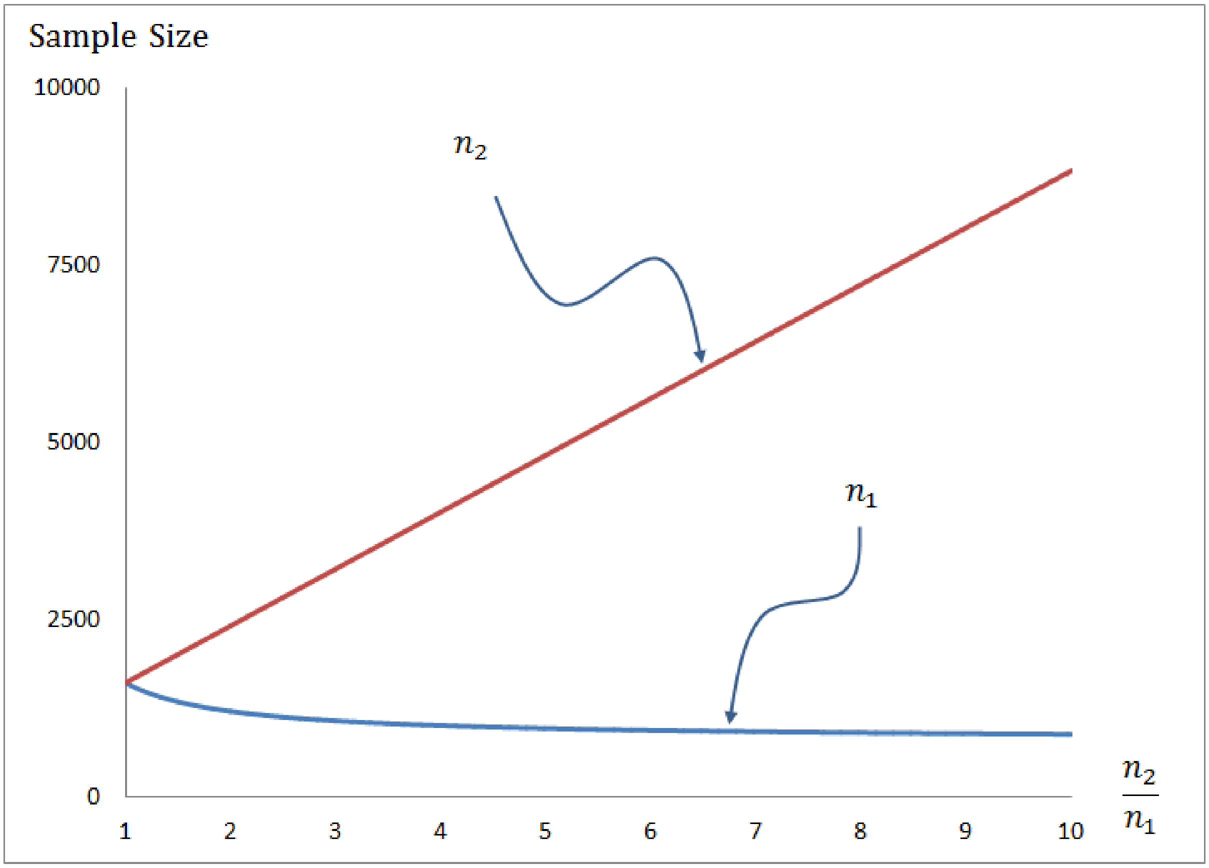


Figure 3.11. An example of feasible relative sample sizes

from population 1 in return for the increases in sample size from population 2. So while there is some scope from "transferring" observations between domains, the marginal gains from doing so (in terms of the smaller sample sizes in one domain) are rapidly diminishing and eventually flat.

We have certainly put the reader through a great deal of math in the this subsection. Fatigue (and perhaps even a wee bit of exasperation) would certainly be understandable. The reason we took the course that we did (as opposed simply to providing the eventual formulas) is that we wanted to convey the logic, rules, and tricks typically applied in generating these relatively straightforward sample size estimation formulas. There are an incredible number of test types, more than the authors could ever anticipate, that can in principle be considered in this framework. And the payoff for the reader is potentially quite high. That is, because the reader understands the logic behind the derivation of sample size formulas for these simple sampling goals, (s)he will likely be able to derive a sample size formula themselves in the event that they are presented with a sampling goal not quite captured by the traditional formulas we have discussed.

3.1.3 Translating Sample Size into Primary Sampling Units

Sample size estimates such as those produced by the formulas introduced in the last subsection produce sample sizes in terms of the units of analysis. However, as we learned in the last Chapter, sampling may not begin with such units. We need somehow to translate the sample sizes produced by the sample size estimation formulas into those units that will actually be selected in the first stage of selection.

Consider, for instance, estimation of the contraceptive expenditures for the population of women ages 15–49 in some nation. Estimates of sample size from the formulas would be in terms of women ages 15–49. However, those sample size estimates are essentially variance based. Consider the two-sided one-sample sample size estimator for a continuous variable:

$$n = \frac{\left[\sigma_Y \cdot (z_{1-\alpha} + z_{1-\beta})\right]^2}{d^2}$$

A key ingredient to the derivation of this formula was the variance

$$V\left(\hat{\overline{Y}}\right) = \frac{\sigma_Y^2}{n}$$

This isn't the variance of $\overline{Y}$ under any circumstances. Instead, it is the sample size based on simple, unstratified, equal-probability independent sampling with replacement.

For present purposes the "with replacement part" is probably less important than the rest. The core issue is that this variance estimator does not assume any of the "selection design" features considered in the last chapter (unequal probability of selection, multistage selection with possible clustering, stratification) occurred.

Recall, however, that we introduced the design effect $deff$ in the last chapter:

$$deff = \frac{V^d}{V^s}$$

where the d means "per whatever design we will use" and the s means simple, independent sample selection without any of those design features. Rearranging, we have

$$V^d = deff \cdot V^s$$

Using this new variance V^d we would have

$$n^{deff} = deff \cdot n = deff \cdot \frac{\left[\sigma_Y \cdot (z_{1-\alpha} + z_{1-\beta})\right]^2}{d^2}$$

where n^{deff} is the design effect adjusted sample size estimate (i.e., the sample size estimate that reflects the likely sampling variance per the design effect).

Thus, step one of going from the output of the sample size estimation formulas is to multiply the result by the design effect $deff$ to get a sample size that reflects the likely consequences of the sampling design that will actually be pursued. But what value should one use for $deff$?

One often hears of reflexive rules of practice for design effect priors such as "we just assume a design effect of 2" or "we just assume a design effect of 1.5." This is a terrible practice.

First, the assumed design effect under reflexive rules is often on the modest side. The temptation to choose a lower prior assumption/estimate for the design effect is a powerful one. It can convince us that we can get away with a more modest sample size than would be necessary with a larger assumed design effect. But unless this lower prior is justified, in the end we are just cheating ourselves.

Moreover, design effects aren't generalizable quantities: they depend on the sampling design and on the indicator. To get some sense of their indicator dependence and variability, in Table 3.2 we report select design effect estimates for the 2007 Bangladesh Demographic and Health Survey (BDHS) from Table B.2 of the 2007 BDHS report (NIPORT, et al., 2009). Demographic and Health Surveys (DHSs) are often a rich source of potential prior information about likely design effects since it is the standard practice in DHS reports to provide sample-based design effect estimates for many indicators in the "Estimates of Sampling Errors" Appendix to the report.[38] Design effects can also be easily computed from publicly available DHS data sets. In STATA, for example, the command `estat effects` can be executed after estimating a parameter of interest to obtain the design effect of a given sample design.

As the reader can plainly see, the design effect varies a lot from indicator to indicator. The

[38] The DHS can be a little bit confusing to interpret, however. For instance, in NIPORT, et al. (2009) they state (p. 246) "the procedure computes the design effect (DEFT) for each estimate, which is defined as the ratio between the standard error using the given sample design and the standard error that would result if a simple random sample had been used." This really is not the design effect as we require for sample size estimation. Rather, it is the design factor. One must square the results in the column for "DEFT" in a standard DHS sampling error table to arrive at what we are looking for sample size estimation purposes, i.e., the ratio between the variance.

range exhibited here certainly is not the limit of what can happen. In the authors' experience we have seen design effect estimates that ranged from well below 1 to astounding values in the low teens. That said, the numbers in Table 3.2 are more typical of the range that we tend to see.

Table 3.2. Design effects
2007 BDHS

No education	3.952144
Currently pregnant	1.488400
Treated with oral rehydration salt	1.408969
BMI < 18.5	2.461761
Has heard of HIV/AIDS	6.538249
Knows about condoms to prevent HIV/AIDS	3.583449
Knows about limiting partners to prevent HIV/AIDS	4.182025

What Table 3.2 also illustrates is that reflexive rules are often unnecessary, as more informed priors are often readily available from earlier work. Thousands of surveys have been done around the world, exploring all sorts of indicators generated from all kinds of estimation approaches and sampling designs. From these we can often get a reasonable ballpark sense of the sort of design effects we are likely to encounter for our own particular application. Even if the precise indicator is not available from a survey of similar design from the same society, a similar one probably is.

One word of caution we would offer is about design effects estimated from small samples, and in several senses (for example, only a few clusters). We have noticed in our own work that small sample estimates of the design effect can often be fairly wild (by which we mean implausible). Indeed, it is not hard to simulate this phenomenon with either an existing sample or made-up data. The takeaway is that the reader should be careful about wild (low or high) design effect estimates formed with small samples. In our experience the design effect estimates take a certain sample size to "settle down," but we are unaware of a whole lot of convincing science about what "certain sample size" means in this context.

We can offer little general guidance worth following regarding any reflexive rules of our own for cases where a decent prior from some earlier survey truly is not available. We have on many occasions been surprised by *ex-post* design effect estimates that defied our intuitive expectation. A few patterns have occurred to us, but only in the roughest sense:

- Design effects are often higher for indicators related to knowledge, attitudes and cultural norms. We theorize that this reflects common culture and media market experience at the cluster level, but cannot be sure. For these kind of indicators it is not unusual to see design effects in the 3–6 range.

- It seems that relatively rare events tend to have lower design effects. An example would be the maternal mortality ratio, which captures the number of maternal deaths per 100,000 live births. The challenge with this indicator is that even in high maternal mortality societies we might not observe many maternal deaths in a typical cluster over a reasonable survey recall window. Maternal mortality, however, also reflects the paradox of some rare events: one would think that their probability would be highly responsive to cluster-level circumstances such as access to health facilities. Thus, we cannot be sure if the small design effects we observe for them are real or reflect a different kind of "small sample" problem in estimating them.

- Relatively common behaviors (for example, modern contraceptive use in a society where modern contraceptive prevalence is around 50 percent) often have design effects that are neither particularly enormous nor tiny. There are often in the 1.8–3 range.

The reader should refer to these "guidelines" only as an act of desperation; although we say that they hold roughly, we have seen many counterexamples as well.

Once the design effect has been accounted for, the main challenge is to translate the design-effect adjusted sample size of analysis units (henceforth just analysis units) into a unit of selection more useful for planning the survey (logistics, budgeting, etc.). Usually this unit is the household or the cluster. In very general terms the steps are:

1. Form an estimate of the number of analysis units expected per respondent household.

2. Divide the estimate of required sample size of analysis units by the number of analysis units expected per respondent household and then round up to the nearest household.

3. Divide the result of 2 by an estimate of household non-response and then round up to the nearest household. The result is the number of households that need to be selected.

4. Divide the number of households by the target number of households to be selected per cluster and round up to the nearest cluster (since it is always awkward and sometimes impossible to select a fraction of a cluster). This is the number of clusters to be selected.

This process can be generalized to several layers of selection above the household. The basic logic would be the same.

Obviously, 1–3 involve estimates of the following sampling priors:

1. The number of analysis units expected per household

2. The household non-response rate

The first prior captures both the demographic realities of households as well as the likely non-response by the analysis units within households. Clearly, sampling priors (i.e., prior assumptions about sampling parameters) are required to be able to do this. A very good place to begin is earlier surveys that considered observed units to be used in your study.

In the simplest terms, from those earlier surveys one can learn how many units of analysis were observed from the number of households that responded (or were selected). In forming an estimate of units of analysis observed per respondent household, one must be careful using official reports on the earlier surveys as the basis for the estimate of this ratio since it can often be unclear whether the reports provide weighted or unweighted numbers. (Since the objective is about actual selection targets, it is the latter that is of interest for sample size estimation.) In the absence of survey evidence on the ratio of analysis units to households, some evidence can possibly be gleaned from censuses or other sources of demographic estimates.

To perhaps make this slightly clearer, we consider two examples. First, suppose that we wish to estimate the modern contraceptive prevalence rate among married women ages 15–49 in some society. Ideally, we would try to form a prior sense of how many married women ages 15–49 to expect per household interviewed or selected from a prior survey that examined contraceptive prevalence. Barring that, one might turn to any survey that sampled households and then interviewed married women within them. The reason the first option is preferred is that different types of survey topics might prompt different non-response rates. One could also look at household rosters even from surveys where no women were interviewed. In the international public health context, there are many, many options by way of earlier surveys that might provide useful guidance, including large-scale, ongoing multi-country survey agendas such as the DHS, Living Standards Measurement Surveys, etc.

Sometimes figuring out the number of respondents per household is a bit trickier. For instance, the authors recently collaborated on developing a sampling plan for a potential maternal mortality survey in Bangladesh to be implemented in 2016. The primary, overarching objective of the survey was to develop a nationally representative estimate of the maternal mortality ratio for Bangladesh for the three-year interval preceding the survey. The maternal mortality ratio is, as we have mentioned, the number of maternal deaths per 100,000 live births. Therefore, the unit of analysis is births.

What we need to figure out is the number of households in which we would need to observe a birth in the study interval (the preceding three years). First, we assume a 2016 crude birth rate of 19 and an average household size of 4.4. These assumptions are based on recent trends observed in the DHS. For instance, preliminary results from the 2014 Bangladesh DHS (NIPORT, Mitra and Associates, & ICF International, 2015) suggest a crude birth rate of 22.2, making 19 a conservative prior estimate for 2016. These assumptions imply that we would expect to observe 0.0836 ($= (19/1000) \cdot 4.4$) births per household per year or 0.2508 per household over the three-year recall window. This implies that we would need to select 3.987 ($\approx 1/0.2508$) households to observe one birth in the preceding three years. So figuring out births per household in the preceding three years, and hence the number of households we would need in order to observe births in the preceding three years, requires a somewhat more elaborate indicator: the crude birth rate. Nonetheless, decent prior estimates for this rate should be available from any number of demographic surveys available in a reasonably recent interval for most countries.

Although there may be many earlier surveys from which to form sampling priors for translating a sample of units of analysis into one of households, it will still not necessarily be easy to do so. For one thing, many of the societies in which we conduct surveys are experiencing rapid demographic, family planning, and health transitions. Thus, in some sense the challenge is rather like making a pass to a teammate in ice hockey — one must pass the puck not to where the teammate is at the instant the pass is initiated, but to where the teammate will be when the puck (hopefully) reaches him or her. This process can be tricky, as it requires the passer to project where his or her teammate will be. So too with generating survey priors. What we are often trying to do is, from the base of the results of an earlier survey, project future circumstances at the time of our own survey for a society in the course of complex, rapid transition.

Recently, one of the authors was tasked with developing some sample size estimates for an impact evaluation in Rwanda that required translating various analysis samples of various cohorts of youth into household sample sizes. The complexity came from the fact that Rwanda had been undergoing a rapid fertility transition which saw the total fertility rate fall very rapidly from 6.1 in the 2005 DHS to 4.6 in the 2010 DHS. This kind of change clearly has implications for within-household demographic patterns. Unfortunately, a more recent DHS-based estimate was not available as of the time of sample size estimation (the 2014–15 DHS results having not been released yet). It was extremely difficult to project where total fertility might be, and this was critical for understanding the likely number of youth that we might observe per household. The author decided to discuss the issue with a colleague with expertise in family planning and the fertility transition. She predicted that progress on fertility had likely slowed in part due to method mix as of 2010. Sure enough, a few weeks later the preliminary results for the 2014–15 DHS were publicly released, with an estimate of an overall total fertility rate of 4.2: a decline, but on the low end of what seemed possible based on trends from 2005–10. Still, having already committed to this fertility projection, there was some apprehension in advance of learning the preliminary 2014–15 DHS numbers.

One final factor for which one may need to adjust is attrition. This is an issue in longitudinal studies. Suppose, for instance, that the sample size estimation was designed to detect a difference in some indicator between populations at the endline of a longitudinal survey. Then the sample size that must be selected toward this goal at baseline must be adjusted for likely attrition from baseline to endline.

The longitudinal data infrastructure in lower and middle income nations is, as a gross generalization, quite sparse. It is thus often rather hard to know how much attrition to expect. The Indonesian Family Life Survey (IFLS) has had very high sample retention rates (in the mid- to upper-90s), but this high rate was thought stunning and praiseworthy at the time of the first follow-up round of the IFLS. The experiences since then have been mixed.

The authors have been involved in and monitored several longitudinal surveys and there are a few clear lessons:

1. The tracking instrument that is used to find respondents post interview is critical and should capture contact and potential tracking information not only about them, but also family, friends, colleagues and other associates.

2. Mid-course (i.e., between panels) tracking (sometimes called “cohort maintenance”) can be very helpful, particularly when panels are more distantly spaced in time.

3. More challenges in tracking should be expected with more mobile, dynamic populations. This generally applies in urban settings but can also be true, for instance, in rural settings with high

rural-to-urban or circular migration (perhaps pursuing seasonal work). Some subpopulations are typically more mobile than others.

4. Be conservative in your estimate of attrition (such as when uncertain, which will usually be the case, assume higher amounts of attrition). Attach some skepticism to claims that "this survey will be different. We won't have high attrition because..." These kind of arguments have been floated at the design phase of nearly every longitudinal survey for which we are aware of design phase discussions. The arguments sometimes prove true. And sometimes they do not prove true.

If we had no useful sense of likely attrition on which to base a figure for sample size estimation, we would probably assume *at least* 20 percent.

3.1.4 On Guesstimation

We are now at a point in the discussion where it should be becoming clear why this chapter has the title that it does. Sample size estimation is the scientific formalization of an educated guess. Even the comparatively straightforward sample size estimation models considered thus far involve assumptions. We assume that certain distributions represent good approximations, for two-sided tests we assume that some terms are essentially 1 or 0, etc.

Once we are past these assumptions we then need to think about the variance consequences of the sampling design we will actually pursue as captured by the design effect. Design effects are estimates, and as such subject to sampling variation. We generally don't know how well real world single-sample estimates of the design effect perform. Frankly, we admit that we (the authors) know little about sampling variation of estimators for the design effect. Indeed, best practice for the practical estimation of the design effect itself is not completely settled at this point.

Once we are done with the design effect adjustment, we need to translate the sample size of analysis units into sampling units that will serve as primary sampling units, be the focal point for thinking about budgeting and planning, etc. For the sort of surveys we tend to perform (which focus on individual or micro-level health and human welfare behaviors and outcomes) this process typically relies on prior estimates of demographic patterns within households. Perhaps these estimates are adjusted to (hopefully) predict accurately circumstances of sample selection and fieldwork.

Finally, we adjust for factors like non-response and attrition, which are impossible to know for sure ahead of time. All we can do is look to past experiences with surveys that in design, execution, and setting are unlikely to match exactly those that we will face.

Good sample size estimation is critical: failure of sample size estimation can mean the failure of surveys, at least in terms of their ostensible goals. But sample size estimation is not exact. It relies on approximations, guess, assumptions, and projections at virtually every stage. It is important to remember this when thinking about one's philosophy of sample size estimation. Above all, we have, and will again now, argue that the uncertainty surrounding sample size estimation is an argument for conservative sample size estimation.

3.1.5 Multiple Sampling Goals

So far our discussion of sample size estimation has been based on the premise that we are concerned with the sample size required to learn about the true value for a single population parameter (we usually call it $\overline{Y}$). Even in the case of comparing estimates from samples from two different populations, we had one population parameter in mind (for example, contraceptive prevalence) that we wished to compare across populations.

But what if we are motivated to perform a survey to learn several different things? This is in fact more often the case than not. For example, we might wish to learn about:

- The contraceptive prevalence rate among married women aged 15–49
- Whether that contraceptive prevalence rate differs between rural and urban areas
- The antenatal care rate for all births in the past 3 years
- Whether the poor have a different antenatal care rate than everyone else
- The percentage of children experiencing malnutrition
- Whether the percentage of children experiencing malnutrition differs between those from households enrolled in some national program and those not enrolled

We could go on and on adding goals. In the program impact evaluation context, we tend to want to learn about program impact, but in terms of what outcomes and for whom? We could in principle gauge impacts in terms of all sorts of behaviors, attitudes and beliefs and outcomes for all sorts of populations: participants, everyone, rural participants, women, the poor, poor participants, etc.

Pursuing multiple sampling goals forces us to compute sample size requirements for the different indicators, populations, and sampling goals (for each indicator and population) involved. The simplest way to do this is to calculate the sample size required for each goal and then determine the maximum of these sample sizes and goals.

Intuitively, let n_{ij} be the required sample size for population i given sampling objective j (where the goal involves some goal for some indicator) for that population. The binding sample size for that population i in the face of J goals is

$$max_{j=1,..,J}\left\{n_{ij}\right\}$$

The point is that, for a given population the required sample size is the maximum required sample size for every indicator and goal indicated for that population.

In practice it is not quite this simple. For one thing, we might be interested in more than one population, making comparison complicated. For instance, it is not really meaningful to say that the sample size required for married women is larger than that required for 6–59 month old children. To really make a comparison, each would need to be translated into a required sample size in terms of some comparable unit, such as a household or cluster.

Consider, for instance, a multistage design whereby we first select clusters, then households, and then, to fix ideas, our two populations of interest within households, children ages 6–69 months and married women ages 15–49 years. It is hard to see why we would opt to select one sample of households to obtain our required sample of women and another to obtain our required sample of children.[39] Instead, we would obtain both from the same sample of households. Thus, the most straightforward way to obtain the sample size of households required is to determine the maximum of the household sizes required for all of the sampling goals involving women and children.

Often, sample size estimation needs to be performed at the strata level. The typical reason for this is a sampling goal defined at the strata level. For instance, we might have sampling goals for rural areas in a sample designed to produce a nationally representative sample as well. We would then determine the "binding" sample size of households at the strata level, and our overall sample size would then be the sum of these binding household sample sizes across strata. If there are cross-strata sampling goals that require a sample size across strata that exceeds the total from the strata level goals, then the cross-strata sampling goal sample size would be binding, and not the totals from the binding goals in each strata.

There are other, more sophisticated ways of going about this process. Some use linear programming, factor in survey costs to find some kind of optimal sample size, etc. While these are quite reasonable and persuasive theoretically, and are used in many surveys in wealthier countries, it has not been our experience that they would be particularly useful for a lot of international work. The discussion of survey goals is typically sufficiently fluid, and transaction costs of determining costs of each possibility covered in the conversation sufficiently high, that more sophisticated methods are less practical. The real lesson of sampling discussions, in our experience, is to conclude them before priorities for the survey can shift!

One final situation that bears mentioning is what to do when one thinks *ex-post* (i.e., even after survey fieldwork is complete) sampling goals might be introduced. Specifically, our concern is with the case where these goals really cannot be anticipated in advance. This possibility is natural with large surveys because they present, for example, the potential opportunity for developing reasonably accurate estimates for subpopulations for which typical survey sizes would be insufficient. As long as the cost of attempting to accommodate these potential future goals is not large, it seems reasonable to try to build into the sampling design the flexibility to do so.

To fix ideas, MEASURE Evaluation was involved in the sampling for the 2010 Bangladesh Maternal Mortality and Health Care Survey (2010BMMS) (NIPORT, MEASURE Evaluation, & icddr,b, 2012). The overarching goal of that survey was to obtain a fairly precise nationally representative estimate of maternal mortality in Bangladesh for the three years preceding the survey. This goal required a huge (roughly 175,000 household) sample size. Actual selection involved fairly involved stratification. First, the largest subnational administrative unit in Bangladesh is the division. Within each division, we selected separately for urban, "other urban," and rural areas. With six divisions, this implied 18 strata.

[39] We cannot say that we would never elect to do this, but certainly in the authors' experience this option has rarely been pursued.

At the time of the design of the sampling for the 2010BMMS, the survey team was aware that, over the long interval within which survey planning and preparation, fieldwork, analysis, and preliminary dissemination of results would occur, the public health policy discussion in Bangladesh would evolve, and with it would evolve the most urgent information needs. Of particular concern was that subnational estimates might be important. Some of these could be anticipated to some degree, but there were two big problems in building sample sizes around them:

1. Specific, precise goals remained illusive.

2. There were not very good prior estimates on which to premise sample size estimation.

What was needed was some way of parsing the national sample across strata that might leave some degree of flexibility to meet with reasonable precision some (if not many) subnational estimation goals that might present themselves in the future.

An obvious candidate would be to divide the sample among the strata according to their share of population size. Thus, if n^O was the overall sample size, the sample size of the i^{th} strata from the 2010BMMS would be

$$n_j = n^O \cdot \frac{pop_j}{\sum_{j=1}^{18} pop_j}$$

where pop_j is the population size of the j^{th} stratum (available from the census, demographic estimates, etc.). By apportioning sample size by strata population share, this strategy has the advantage that it should not necessarily by itself contribute significantly to weight variation (though it likely would contribute *some* to weight variation because pop_j will likely prove to have been somewhat inaccurate in practice).

The potential problem with this approach is that it might leave a fairly small sample to the smallest strata. In fact, the largest stratum was around 84 times the size of the smallest one by population. By simple apportionment by sample sizes, the smallest stratum would have received only around 0.3 percent of the 175,000 or so households for a final sample size in that stratum of around 500 households (probably not enough to satisfy many estimation goals).

Instead, we apportioned by the square root of population size. The sample from the j^{th} stratum is then[40]

$$n_j = n^O \cdot \frac{\sqrt{pop_j}}{\sum_{j=1}^{18} \sqrt{pop_j}}$$

This approach left the smallest stratum in the 2010BMMS with a sample size of around 2,700 households, not huge but likely enough to have some flexibility to meet many estimation goals. To be sure, this approach may exacerbate weight variation somewhat, but we felt it struck a decent balance between that concern and the need to retain flexibility for subnational sampling goals.

[40] In practice, we first apportioned across divisions according to their share in the square of population sizes across divisions, and then across the urban, rural and other urban strata within each division according to the share of the square of population size of each division. This two-stage parsing is not necessarily preferred or inevitable, but was instead driven simply by the sequence in which population numbers for the divisions and strata became available and the need to deliver sample size estimates by division early in the process for contractual needs (it was division-level sample sizes that were essential for contracts for the survey firms involved in the 2010BMMS).

3.1.6 Finite Populations

We conclude this section with a topic that periodically pops up when estimating sample size for program impact evaluations: whether to apply a finite population correction to the variance used for sample size estimation. The finite population correction is an adjustment that is offered to the sampling variance when the sample is selected without replacement. The typical effect of the finite population correction is to yield a final sampling variation that is smaller than the standard, or "infinite," population variance.

To begin with, let us first digress to talk briefly about a topic that inevitably has to be addressed in any sort of sampling manual: the method of sample selection. In particular, we distinguish sampling *with* replacement from sampling *without* replacement. In the former case, when we select a sampling unit from a population, it is not removed from eligibility for further selection as sample selection continues. In other words, a given sampling unit can be selected more than once. In the latter case, a sampling unit is no longer eligible for selection once it has been selected.

Sampling without replacement generally involves smaller sampling variance for estimates of, say, population means with the resulting samples. The reason for this is pretty simple: there are more possible sample combinations that can emerge under sampling with replacement than sampling without replacement. If the population has N sampling units and we are selecting a sample of size n, then when sampling with replacement the number of possible samples we could draw is

$$N^n$$

By contrast, when sampling without replacement, the number of possible sample combinations (disregarding order of selection for particular sampling units) is

$$\frac{N!}{n! \cdot (N-n)!}$$

Since $n \leq N$,

$$N^n > \frac{N!}{(N-n)!}$$

Meanwhile, $n!$ is greater than 1. Therefore,

$$N^n > \frac{N!}{n! \cdot (N-n)!}$$

In other words, more possible samples (in terms of the combinations of sampling units involved) could result from sampling with replacement than sampling without replacement. This feeds the greater sampling variation to estimates that one sees (at least in theory) with sampling with replacement.[41]

[41] In some sense the distinction between with and without replacement selection, and its implications for variance, is a touch silly for practical purposes: in *practice* we typically select samples by systematic random sampling for which, technically speaking, there is no theoretically valid formulaic estimator of the variance. In response to this we tend to cheat somewhat by using the without-replacement formula. Nonetheless, it is worth bearing in mind that concern

Thus, for any given population size, with-replacement sampling leads to more different potential samples and therefore, somewhat crudely, more variation. Notice as well that as the sample size n gets large relative to the population N, under sampling without replacement we are beginning to run out of possible combinations of samples. At the limit, there is only one possible sample when $n = N$, N possible samples when $N - n = 1$, etc.[42] The point is that in the without-replacement case, potential sampling variation becomes increasingly limited as n/N approaches 1.

At the same time, most conventional formulas for variance used in sample size estimation to this point, such as

$$\frac{\sigma^2}{n}$$

were derived under an assumption of selection from an "infinite" sample. One implication of assuming an "infinite" sample is that changes in the sample size n really did not meaningfully affect the potential sampling variation, regardless of whether selection was with or without replacement. However, when sampling without replacement from a less than infinite population of size N this is no longer true. Increases in the sample size n begin to act as a drag on potential sampling variation in the sense discussed at the beginning of this paragraph.

This suggests that we might be overestimating variance with such conventional formulas when we select without replacement. For the without-replacement selection case, we thus need to make an adjustment to the variance under the conventional formulas. Specifically, we need to multiply estimates of the variance of an estimate when selecting without replacement samples of size n from a finite population of size N by a finite population correction factor (FPC)[43]

$$FPC = \frac{N-n}{N} = 1 - \frac{n}{N}$$

In truth, if n is really small compared with N, the finite population correction

$$1 - \frac{n}{N}$$

is essentially equal to 1. Conventional thinking usually suggests that once the sampling fraction

$$\frac{n}{N}$$

gets above 0.05 the FPC becomes important. (If $n/N = 0.05$ the FPC equals 0.95.) That said,

with the finer distinctions between these two approaches must be balanced against the fact that in practice we will probably use a selection method to which neither fully speaks and to which we technically (that is, theoretically) cannot apply the variance formulas from either.

[42] Notice that there are far more possible samples under sampling with replacement when $N - n = 1$. Indeed, as sample size n increases, the number of potential sample combinations actually *grows* under with replacement sampling.

[43] Technically speaking, $FPC = \frac{N-n}{N} \cdot \frac{N}{N-1} = \left(1 - \frac{n}{N}\right) \cdot \left(\frac{N}{N-1}\right)$, which leads to a commonly adopted notation of $\frac{N-n}{N} \cdot \frac{S^2}{n}$ (where $S^2 = \frac{N}{N-1}\sigma^2$) for the sampling variance under sampling without replacement. It is mathematically derived from a correlation in selection probabilities between two units due to sampling without replacement. The mathematical derivation is beyond the scope of this manual and an interested reader is referred to Cochran (1977) and Lohr (1999). Notice that for large N, $\frac{N}{N-1} \approx 1$, leading to our simplified notation $FPC = \frac{N-n}{N}$.

there really is not a convincing point where one can say that the FPC matters — as long as $n < N$ the FPC is greater than 0 but less than 1.

Notice, however, that if $n = N$ the FPC equals 0, and thus the sampling variance of the estimate is zero. This gets to an important point (one we will return to below) about the finite population approach. If you observe everyone in the population, then you are not estimating but instead calculating exactly the population parameter. In other words, because we observe the entire population we can know exactly (i.e with certainty) the population parameter for them. Estimation, by contrast, is about learning something about a population from a partial observation basis.

This gets to an important point about the estimand from the vantage point of the finite population approach — it is defined by the actual experiences of the population in question at the point of observation. A "superpopulation" approach, by comparison, is focused on the underlying true process behind what is observed (i.e., the distributional process that shapes outcomes). Hence, from the superpopulation perspective, even if we observed the entire population at a point in time we could estimate a parameter but not compute it.

These distinctions were glossed over in the simulated example used to discuss sample selection in the last chapter. Then the focus was on a population of 5,000 households and the oft-stated goal of learning about "the true population average daily expenditures of 39.29789." By this we meant that, were we able to observe all 5,000 households at the point of observation for which we were selecting, we could have computed an average household expenditures across them of 39.29789 because this was the actual average at that point in time for the 5,000 households. In other words, the estimand was what actually happened in the full population at the point of observation, as opposed to some underlying process that generated those 5,000 outcomes at that point in time. In other words, in the last chapter the numerical example was based on the finite population perspective. We essentially glossed over the difference because the focus there was on understanding the consequences of different selection approaches, and to delve too deeply into the finite/infinite population perspective would have muddied the waters for too little gain.[44]

We now consider the consequences of the finite population correction for sample size. We use the case of a one-sided, one-sample test. Recall that determining the sample size came down to the condition

$$-z_{1-\beta} = z_{1-\alpha} - \frac{d}{\sqrt{V\left(\hat{\overline{Y}}\right)}}$$

Using the variance

$$V\left(\hat{\overline{Y}}\right) = \frac{\sigma_Y^2}{n}$$

we arrived at the sample size estimate

$$n = \left[\frac{\sigma_Y \cdot (z_{1-\alpha} + z_{1-\beta})}{d}\right]^2$$

[44] That being said, however, the reader may wonder whether the FPC should then have been employed. In the case of some of the larger samples (for example of $n = 750$, for which the FPC would be .85) the answer is probably yes.

We can think of this as being akin to the sample size required for selection from an "infinite" population.

In the finite sample size case, we begin with the variance

$$V\left(\hat{\overline{Y}}\right) = \left(\frac{N-n}{N}\right) \cdot \frac{S_Y^2}{n}$$

We thus begin with the condition

$$-z_{1-\beta} = z_{1-\alpha} - \frac{d}{\sqrt{\left(\frac{N-n}{N}\right) \cdot \frac{S_Y^2}{n}}}$$

This leads to

$$\frac{1}{\sqrt{\left(\frac{N-n}{N}\right) \cdot \frac{S_Y^2}{n}}} = \frac{z_{1-\alpha} + z_{1-\beta}}{d}$$

and then

$$\frac{N \cdot n}{N-n} = \left(\frac{S_Y \cdot (z_{1-\alpha} + z_{1-\beta})}{d}\right)^2$$

Rearranging further we have

$$n = \frac{1}{N} \cdot \left(\frac{S_Y \cdot (z_{1-\alpha} + z_{1-\beta})}{d}\right)^2 \cdot (N-n)$$

or equivalently

$$n \cdot \left[1 + \frac{1}{N} \cdot \left(\frac{S_Y \left(z_{1-\alpha} + z_{1-\beta}\right)}{d}\right)^2\right] = \left(\frac{S_Y \cdot (z_{1-\alpha} + z_{1-\beta})}{d}\right)^2$$

Rearranging and solving, we have

$$n = \frac{\left(\frac{S_Y \cdot \left(z_{1-\alpha} + z_{1-\beta}\right)}{d}\right)^2}{1 + \frac{1}{N} \cdot \left(\frac{S_Y \cdot \left(z_{1-\alpha} + z_{1-\beta}\right)}{d}\right)^2}$$

$$\approx \frac{\left(\frac{\sigma_Y \cdot \left(z_{1-\alpha} + z_{1-\beta}\right)}{d}\right)^2}{1 + \frac{1}{N} \cdot \left(\frac{\sigma_Y \cdot \left(z_{1-\alpha} + z_{1-\beta}\right)}{d}\right)^2}$$

for a large N (i.e., $N - 1 \approx N$) so that $S_Y \approx \sigma_Y$. This is somewhat akin to the original sample size estimator divided by something that is in some sense one plus the sampling fraction.

It should be clear that, since

$$0 < \frac{1}{N} \cdot \left(\frac{S_Y \cdot (z_{1-\alpha} + z_{1-\beta})}{d}\right)^2 \leq 1$$

the sample size for the finite population case will be less than the sample size when we do not consider a finite population correction. Indeed, it could be reduced up to 50 percent from the figure that would have emerged without consideration of a finite population correction. It is not uncommon for us to hear someone advocate for considering a finite population correction largely because it leads to smaller sample sizes, other things being equal.

The finite population approach to variance estimation is thus quite an attractive possibility. However, it is not necessarily an appropriate one even when the population is finite (i.e., even when n/N is large[45]). The general tenor of the discussion in mainstream statistics would instead seem to place the emphasis on the goals of estimation.

Most importantly, the consensus in mainstream statistics would seem to be that finite population corrections are generally inappropriate when the goal of a survey is to test a hypothesis. For instance, in one of the classic statements on the matter, Cochran (1977) writes in the context of comparison of domain means:

> It is seldom of scientific interest to ask whether [*means from two populations are exactly equal*],[46] because these means would not be exactly equal in a finite population, except by a very rare chance, even if the data in both domains were drawn at random from the same infinite superpopulation. Instead, we test the null hypothesis that the two domains were drawn from *infinite* populations having the same mean. (p. 39)

(See as well, for instance, Deming, 1966; Korn & Graubard, 1999; etc.). Generalizing this, the null hypothesis is not likely to hold exactly in a finite population or any sample from it, even if the finite population itself is a random subset of some superpopulation for which the null hypothesis does indeed hold exactly.

On the most basic level, casual estimation and impact evaluation can be viewed as an exercise in hypothesis testing. Essentially, the notional null hypothesis is "no impact." A manifestation of this that we might look for at the most elementary level is differences in average outcomes between participants and non-participants. To fix ideas and sweep aside many complications, imagine that our sample emerged from a successful experiment and thus the sample (and population, where our particular interest for the present purposes and hence the focus is on a *finite* population), mean outcomes for participants are representative of what the average person in the population would experience in the participation state and the mean outcomes for non-participants are similarly representative. We would never expect these to be exactly equal in any given samples or finite populations (even if the program in truth had no impact on behavior). We reject the null not if the sample mean outcomes differ between our samples of participants and non-participants (since this is virtually guaranteed to occur), but instead if they diverge enough to conclude that in fact the program must have some impact.

We have sometimes heard it argued that finite population corrections to variance can be justified if the hypothesis is stated explicitly in "finite population terms." As long as $n < N$, one could

[45] When N is explicitly infinite then there is no possibility for the finite population approach and when n/N is small it is not important since little is gained in the way of variance reduction by including it.

[46] Where the italics are our own and designed simply to avoid notational confusion.

in some sense still convince oneself of the appropriateness of proceeding with classical testing of parameters estimated over that sample: owing to sampling variation, it is reasonable to argue that a classical test might lead us to fail to reject the null hypothesis even if the estimates from the sample did not conform precisely with it (they would need only to be *likely* under the null hypothesis and the assumed distribution of the test statistic).

The strangeness of the prospect of a finite population hypothesis becomes more apparent when one considers the prospect of obtaining a value for the parameter(s) of interest from a census (that is, when $n = N$) of that finite population. In that case, a classical probabilistic test (where one fails to reject the null hypothesis if the parameter value calculated is likely under the null hypothesis and some assumed test statistic value) using the parameter calculated across the finite population makes no sense. Rather, the calculated value of the parameter would need to conform exactly with the null hypothesis in order to fail to reject it (and in that instance, we would probably use the stronger language of "accepting" it). Another manifestation of this is that the variance for the calculation (notice we are no longer saying *estimate*) of the parameter across the finite population would be zero. Because the calculation across all N members of the population captures the entire finite population, sampling variation is no longer a meaningful concept.

The trouble is that we know going into this exercise that in most circumstances there is almost no chance that the calculation of the parameter across the finite population will yield values exactly consistent with the null hypothesis. And even in the rare cases where it did, it would be difficult to know what to make of the result.[47] One way of thinking about this is that even when we observe the entire finite population, the information is still based on the experiences of individuals that can be seen as something of a random draw from a distribution of possibilities centered on the truth. In other words, both in practical and conceptual terms the finite population hypothesis test is a strange bird that quickly leads us back to concerns along the lines of Cochran (1977).

Some who question finite population corrections in a setting such as impact evaluation focus on an argument against them that is a kind of subtle variation on this basic critique. To introduce it, we delve a bit more explicitly and deeply into the arena of program impact evaluation. We start by motivating a model for the data generating process behind what we will observe in the sample we eventually produce. We begin with the classic potential outcomes framework (Rosenbaum & Rubin, 1984). Under this approach, all individuals have, with respect to any program, two potential outcomes for each behavior or outcome Y relevant to them: Y^1, which is what they would experience were they to enroll or participate in the program, and Y^0, which is what they would experience if they did not participate.

Let P be an indicator variable that equals 1 if an individual is exposed to or participates in a program and 0 otherwise. This is *observed* program participation. It tells us whether the person actually participated in the program or not.

[47] In this regard, there is something of an analogy to testing whether a broken watch is correct by a single comparison of its reported time with that from some other timepiece. It is a test the broken watch would pass only if that test was conducted during the 2 (out of 86,400) seconds of the day that correspond to the time to which the arms of the broken watch were stuck (and even then the conclusion one would draw would be wrong!).

For each individual we can define their observed outcome Y as

$$Y = (1 - P) \cdot Y^0 + P \cdot Y^1$$

This simply says that the outcome we actually observe for each person depends on whether they participated in the program. If they did, we observe their outcome as a participant in the program (Y^1). If not, we observe their outcome in the absence of participation (Y^0).

We are now in a position to start developing a model for potential outcomes, observed outcomes and program participation. To begin with, the survey observation interval will be t. We consider the following potential outcome and cost equations:

$$Y_{it}^0 = \beta_0 + \beta_2 \cdot X_{1it} + \beta_3 \cdot X_{2i} + \beta_4 \cdot \mu_i + \epsilon_{it}^Y$$

$$Y_{it}^1 = \beta_0 + \beta_1 + \beta_2 \cdot X_{1it} + \beta_3 \cdot X_{2i} + \beta_4 \cdot \mu_i + \epsilon_{it}^Y$$

where X_{1it} is a time-varying observed individual characteristic, X_{2i} is a fixed observed individual characteristic, μ_i is a fixed unobserved individual characteristic, and ϵ_{it}^Y is time-varying unobservable.

The cost of participation is given by

$$C_{it} = \gamma_0 + \gamma_1 \cdot X_{1it} + \gamma_2 \cdot X_{2i} + \gamma_3 \cdot Z_{it} + \gamma_4 \cdot (1 - Z_{it}) + \gamma_5 \cdot \mu_i + \epsilon_{it}^C$$

At time period t individual i chooses to participate (i.e., P_{it} will equal 1) if

$$Y_{it}^1 - Y_{it}^0 - C_{it} > 0$$

or

$$\beta_1 - \gamma_0 - \gamma_1 \cdot X_{1it} - \gamma_2 \cdot X_{2i} - \gamma_3 \cdot Z_{it} - \gamma_4 \cdot (1 - Z_{it}) - \gamma_5 \cdot \mu_i - \epsilon_{it}^C > 0$$

Thus folks will participate if they do better by doing so.

Let us now derive a model for the observed outcome. The observed outcome is

$$Y_{it} = P_{it} \cdot Y_{it}^1 + (1 - P_{it}) \cdot Y_{it}^0$$

$$= P_{it} \cdot \left(\beta_0 + \beta_1 + \beta_2 \cdot X_{1it} + \beta_3 \cdot X_{2i} + \beta_4 \cdot \mu_i + \epsilon_{it}^Y\right)$$

$$+ (1 - P_{it}) \cdot \left(\beta_0 + \beta_2 \cdot X_{1it} + \beta_3 \cdot X_{2i} + \beta_4 \cdot \mu_i + \epsilon_{it}^Y\right)$$

$$= \beta_0 + \beta_1 \cdot P_{it} + \beta_2 \cdot X_{1it} + \beta_3 \cdot X_{2i} + \beta_4 \cdot \mu_i + \epsilon_{it}^Y$$

This is the process behind the observed outcomes we would see from samples of participants and non-participants. This basic model could be generalized in any number of directions to include various types of exogenous observable determinants of cost (variables like Z_{it}), various error component structures, observed and unobserved determinants of the costs and potential outcomes, etc.

From this basic framework we could motivate all sorts of possible program impact estimation data-generating processes and justify all sorts of estimation frameworks, including:

1. A experiment whereby participation is fully randomized
2. Selection on observables
3. Within models
4. Instrumental variables
5. Regression discontinuity design

These various possibilities would be motivated primarily by different assumptions about the γs and βs. And estimation would be aimed at recovering the βs, particularly β_1.

A given individual's observed outcome (and potential outcomes) at time t would not be their "intrinsic" value for Y, Y^1 or Y^0. Rather, it would be the outcome of a stochastic process that depended in part, for instance, on the draws for the ϵs that they received at observation interval t.

This model is not necessarily the only way to characterize the process giving rise to potential and thus observed outcomes. One can certainly elaborate on or improve on it, but what those innovations would have in common is that they would still represent *models* of the underlying process behind outcomes.

Thus, what we are really doing when we use samples of program participants and non-participants to estimate program impact is implicitly not recovering "the" program impact across either those samples or the populations (finite or infinite) that they represent. Rather, we are really estimating impact per an underlying super-population process that generates observed outcomes and participation, regardless of the actual size of the population or our samples from it.

Looked at from this standpoint, in general it makes very little sense to think of program impact evaluation as a finite sample exercise. It is much more persuasive to think of it as recovery of the parameters driven by or behind a super-population process. This "underlying model" argument has a fairly long pedigree in statistics that continues to the present (for example Valliant, et al., 2013).

The upshot of either variant of the argument (the hypothesis-testing approach or the extension to the idea of a super-population process) would seem to be that finite population corrections should be avoided in the program impact evaluation context. It is not an appropriate approach to hypothesis tests or model-driven estimation, both of which apply in the context of impact evaluation.

All of this having been said, this is not an entirely settled matter. For instance, a rather slick approach to the basic idea of the finite population inference has recently been offered in the impact evaluation context. Abadie, Athey, Imbens, and Wooldridge (2014) more or less sidestep the largely philosophical approach to the question of the finite population corrections that has been the focus of our discussion thus far. Instead, they embrace a somewhat different philosophical departure point less motivated in the goals of the estimation:

> The general perspective we take is that statistics is fundamentally about drawing inferences with incomplete data. If the researcher sees all relevant data, there is no need for inference, since any question can be answered by simply doing calculations on the data. (p. 2)

They avoid the super-population argument and basically assume that the actual outcomes for a population at a point of observation define the value for the population parameter. They then bore in on causal modelling and derive variance estimators based on the inherent partial observation challenge in answering counterfactual casual questions (such as that you cannot observe an individual at a point in time in both the participant and non-participant state). The importance of Abadie, et al. (2014) for the present purposes is that it shows, first, that there is not universal agreement about the essentially pessimistic approach to finite population corrections in the setting of hypothesis testing, and second, that there can be creative ways of thinking about this question in the program impact evaluation context.

At this point we, the authors, tend to hew toward the views of Cochran (1977), Valliant, et al. (2013), etc. First, the reasoning behind the super-population approach is reasonable and a natural fit for the impact evaluation setting. Second, the variance estimates without a finite population correction can be viewed as conservative (compared with those embodying the correction), and as the reader has already come to recognize, we tend to be conservative in our approach to sample size estimation. Nonetheless, the line of research opened by work such as Abadie, et al. (2014) is an exciting one that may lead us to alter our approach in future, at least in some circumstance. For now, however, it is our practice generally to eschew finite population corrections in sample size estimation for impact evaluations.

3.2 Randomized Controlled Trials

In this subsection we discuss sample size estimation for RCTs. We tend to think of four main traditions of impact evaluation that more or less encompass most of the practice of impact evaluation (at least estimation-based evaluation with samples of units of observation): RCTs; selection on observables methods (such as regression or propensity score estimation); within models (fixed effects, first differences, difference-in-differences, etc.); and instrumental variables (a category within which we can for present purposes subsume “fuzzy” regression discontinuity designs). Having laid a foundation for thinking about sample size estimation with the classic “work horse” formulas discussed in the last section, we devote the remainder of this chapter to discussing sample size estimation approaches more specific to these four classic approaches. Where there are compelling formulas for sample size estimation, we provide them. Even when there are not, we offer some narrative around the sample size questions and challenges for different impact evaluation approaches. A general theme of the remainder of this chapter is that the options for sample size estimation are limited, and even when they are not the available estimators generally rely on strong assumptions.

We start with RCTs in part for the same reason we did in MEASURE Evaluation’s recent impact evaluation manual (Lance, Guilkey, et al., 2014). Regardless of the limitations of RCTs,[48] they provide a very useful framework for thinking about the notional counterfactual experiment conceptually behind most applications of impact evaluation estimation strategies. We therefore start with them, in the hope that they will provide a departure point that makes the other impact evaluation estimation techniques, and their identifying and behavioral assumptions, a bit clearer.

To begin with, we need to define some terms. We define an experimental sample as one in which program participation is fully randomized. Typically, we think of this as emerging from an RCT since the purposeful choices of programs (in their enrollment criteria, program targeting and placement, etc.) and individuals (in their participation decisions) under real world circumstances would rarely lead to fully random determination of participation status. However, there have been instances where the argument that randomized participation has effectively arisen under real-world circumstances (see Miller, 2010 for one particularly humorous example). When we refer to RCTs in the discussion to follow, we are motivated by the possibility of randomization emerging as an explicit design choice in a planned evaluation.

In this, we can draw a contrast with non-experimental samples, which are samples where assignment to participation status is less than fully random. This can happen for any number of reasons. Failure of full randomization could be the result of an RCT with less than full compliance (for any number of reasons) with assignment to participation status. More generally, in the messy laboratory of typical real world program operations, individuals sort themselves into participants and non-participants, perhaps influenced as they do so by the incentives, rules, and restrictions put before them by the program itself, the government and others as well as their own preferences, constraints, characteristics, and production processes for various human welfare outcomes (such as health).

We again begin with the basic notation of the potential outcomes approach to program impact (Rosenbaum & Rubin, 1984).[49] Each individual i from the population has potential outcomes

$$\left\{Y_i^1, Y_i^0\right\}$$

Y_i^1 is the outcome (which could be a behavior, an actual outcome that captures some channel of health or some other dimension of human welfare, etc.) that individual i experiences if (s)he participates in the program. Y_i^0 is the outcome that individual i experiences if (s)he does not participate. A given individual i's impact would then be

$$I_i = Y_i^1 - Y_i^0$$

This difference captures the consequences of participation for individual i. This is the causal impact of the program for individual i, and is the foundation of our impact estimands for populations of

[48] See, for instance, Heckman and Smith (1995), Deaton (2010), the recent "Worm Wars" debate (Aiken, Davey, Hargreaves, & Hayes, 2015; Clemens & Sandefur, 2015; Davey, Aiken, Hayes, & Hargreaves, 2015; Ozler, 2015;), etc.

[49] Compared with our first presentation of the potential outcomes framework in the preceding discussion of finite populations, we for now eschew the time dimension of the model since it will at this point add notational complication for no additional expositional gain.

interest.

The participation status of individual i is captured by

$$P_i = \begin{cases} 1 & \text{if individual } i \text{ does participate in the program} \\ 0 & \text{if individual } i \text{ does not participate in the program} \end{cases}$$

This is his/her *actual* (as opposed to *potential*) participation status. In real world observational terms, the *observed* outcome Y_i for each individual is

$$Y_i = P_i \cdot Y_i^1 + (1 - P_i) \cdot Y_i^0$$

Thus, we actually observe Y_i^1 if individual i is a participant and Y_i^0 otherwise. Participation status determines which potential outcome we observe for each individual.

Since impact at the individual level

$$I_i = Y_i^1 - Y_i^0$$

is the foundation of most of our conceivable population parameters of interest related to impact evaluation, ideally we would observe impact at the individual level and estimate program impact with representative samples of observations from which we can estimate individual impact at the individual level. Thus, presented with a sample of n observations from some population of interest, we could estimate impact as

$$\overline{I} = \frac{\sum_{i=1}^{n} I_i}{n}$$

This would in some sense be the “purest” theoretical estimate of impact.

Unfortunately, we can’t observe both Y_i^1 *and* Y_i^0 for any given individual i. As

$$Y_i = P_i \cdot Y_i^1 + (1 - P_i) \cdot Y_i^0$$

makes plain, we observe Y_i^1 only if individual i is a participant and we observe Y_i^0 only if individual i is a non-participant. This is the fundamental identification problem of program impact evaluation.

Program impact evaluation in practice thus usually involves estimation of impact based on some sort of comparison of average outcomes between participants and non-participants. Under the basic approach, we compare outcomes by estimating the “group mean” outcomes for participants and non-participants, with their difference then forming an estimate of impact. Interest, at least for estimation purposes, is then with forming unbiased estimates for the population of interest of the mean outcome under participation ($\overline{Y}^1$) and in the absence of participation ($\overline{Y}^0$)[50] under some sort of stochastic process. For instance, though we have for the moment eschewed consideration of time for expositional simplicity, one could still think of $\{Y_i^1, Y_i^0\}$ as representing individual i’s draw for

[50] As an aside, one could reasonably ask how this approach, or even the individual-level approach to impact, lends itself better to the finite or infinite population approach. We would argue that the latter continues to make more sense. Individuals receive a draw $\{Y_i^1, Y_i^0\}$.

potential outcomes for the observational interval.

Suppose that we have a sample of size n from the population of interest. A simple (and obvious) estimator of average program impact $\overline{I}$ based on group mean outcomes for participants and non-participants would be

$$\hat{\overline{I}} = \frac{\sum_{i=1}^{n} P_i \cdot Y_i}{\sum_{i=1}^{n} P_i} - \frac{\sum_{i=1}^{n} (1 - P_i) \cdot Y_i}{\sum_{i=1}^{n} (1 - P_i)}$$

$$= \frac{\sum_{i=1}^{n} P_i \cdot Y_i^1}{\sum_{i=1}^{n} P_i} - \frac{\sum_{i=1}^{n} (1 - P_i) \cdot Y_i^0}{\sum_{i=1}^{n} (1 - P_i)}$$

where $\sum_{i=1}^{n} P_i$ is the number of observed participants and $\sum_{i=1}^{n} (1 - P_i)$ is the number of observed non-participants.

This is essentially the simplest estimator of program impact using samples of participants and non-participants: it estimates impact as the difference in the average outcomes for participants and non-participants. In doing so, it essentially assumes that all of this difference can be ascribed to the experience of program participation. This would be problematic if participants and non-participants differed by more than just the experience of program participation.

This is the motivation for the experimental samples yielded by RCTs. By randomizing participation status, an RCT insures that participants differ from non-participants only in terms of program participation. Therefore:

1. The expected experiences of participants reflect what non-participants would have experienced had they participated: $E(Y^1|P = 1) = E(Y^1|P = 0)$.

2. The expected experiences of non-participants reflect what participants would have experienced had they not participated: $E(Y^0|P = 0) = E(Y^0|P = 1)$.

3. Combining these insights, program participation status reveals nothing about expected values for the potential outcomes: $E(Y^1|P) = E(Y^1)$ and $E(Y^0|P) = E(Y^0)$.

Given these, the simple comparison of mean outcomes between the participants and non-participants

$$\frac{\sum_{i=1}^{n} P_i \cdot Y_i^1}{\sum_{i=1}^{n} P_i} - \frac{\sum_{i=1}^{n} (1 - P_i) \cdot Y_i^0}{\sum_{i=1}^{n} (1 - P_i)}$$

from an experimental sample is an unbiased estimator of average program impact. In its structure, this is absolutely no different from the challenge of estimating the difference in outcomes between two populations discussed in the last section.

Therefore, sample size estimation for RCTs could be based on the basic sample size estimation strategies that we have just discussed. Formally, sample size estimation is typically based on the goal of being able to estimate average program impact of some magnitude d with power $1 - \beta$ and significance α. The hypotheses are typically

$$H_0 : \overline{I} = 0$$

$$H_a : \overline{I} = |d|$$

With this familiar framework applied to the present context, no further discussion of sample size estimation for RCTs is necessary given the ground covered in the last section.

Sample size analysis then proceeds as in the last section. For the case where subjects are individually randomized into program and control groups, the formulas for detecting a difference of d with significance α and power $1-\beta$ are, for the case of a continuous outcome of interest

$$n = 2 \cdot \left(\left[\frac{\sigma_Y \cdot \left(z_{1-\alpha/2} + z_{1-\beta} \right)}{d} \right]^2 \right)$$

while for the case of an outcome in the form of a rate or percentage it would be

$$n = \left[\frac{z_{1-\alpha/2} \cdot \sqrt{2 \cdot \overline{\overline{Y}} \cdot \left(1 - \overline{\overline{Y}}\right)} + z_{1-\beta} \cdot \sqrt{\overline{Y}_1 \cdot \left(1 - \overline{Y}_1\right) + \overline{Y}_2 \cdot \left(1 - \overline{Y}_2\right)}}{d} \right]^2$$

These are the straightforward formulas of the last section.[51]

One note of caution must be sounded about the term σ_Y, both in terms of forming a per-study prior for sample size estimation as well as *ex-post* testing for any differences with the samples of participants and non-participants. Deaton (2010) lists a number of potential wrinkles when applying conventional standard error estimation practices to the difference in the two average outcomes. Most revolve around the possibility of unequal variances among participants and non-participants. As Deaton points out, assuming that the experiment does not introduce a difference in variance between participants and non-participants violates the "minimalist" spirit of RCTs. In principle, if there was a difference in variance between the two groups we have already seen the appropriate formula for continuous outcomes

$$n = \left(\sigma^2_{Y_1} + \sigma^2_{Y_2} \right) \cdot \left(\left[\frac{\left(z_{1-\alpha/2} + z_{1-\beta} \right)}{d} \right]^2 \right)$$

The sticky wicket in practice is that it might be very hard to form useful priors about $\sigma^2_{Y_1}$ and $\sigma^2_{Y_2}$. One possibility would be to appeal to existing representative samples from the population of interest or closely related populations of interest to examine how the variance differs across subpopulations with (statistically significantly) different values for the outcome. Differences in estimated values for σ^2_Y across subgroups with statistically significant differences in outcomes the magnitude of which is in the neighborhood of d might provide some guidance regarding possible priors for $\sigma^2_{Y_1}$ and $\sigma^2_{Y_2}$.

[51] In both of these cases, we assume that there is no correlation of outcomes between those assigned to the two arms of the experiment, meaning that ρ from the term $(1-\rho)$ is equal to 0 and the term $(1-\rho)$ equals 1. If one wishes to accommodate such a possibility of correlation, simply multiple either n by $(1-\rho)$.

A conservative approach might be to use the sample size formula

$$n = 2 \cdot \left(\left[\frac{\sigma_Y \cdot \left(z_{1-\alpha/2} + z_{1-\beta} \right)}{d} \right]^2 \right)$$

using an assumed value for σ_Y on the higher end of those suggested by recent representative samples for the population of interest or closely related populations of interest.

One issue we have glossed over so far is the actual relationship between the processes of randomization and sample selection. This is in part because there are a number of combinations of ways the two processes can occur. However, understanding the roles that they play is essential for making good decisions about study design and will allow us to frame much of the discussion of what follows. It is probably helpful to start by defining terms. First, the population of interest is the population for which we wish to know program impact. Second, the participants are those assigned to participate in the program per the random assignment process of the randomized controlled trial. The non-participants are those assigned not to participate in the program per the random assignment process of the randomized controlled trial. The participants in the RCT are those subject per the RCT to random assignment to participant and non-participant status. We assume that the efforts of those conducting the randomized controlled trial and those conducting a survey to evaluate the program are aligned so that they have the same population of interest in mind.[52]

The goal of randomization is to insure that participants differ from non-participants only by the experience of program participation (i.e., they do not differ in terms of their average pre-participation characteristics). Any deliberate participation assignment mechanism that delivers this state of affairs can be classified as an RCT.[53] By insuring this condition, randomization allows any differences in post-participation[54] average outcomes between participants and non-participants to be ascribed to the experience of participation. Assuming that the individuals who participate in the RCT are representative of the population of interest (an assumption we effectively make in the preceding paragraph), randomization also insures that the participants and non-participants are representative of that population.

The ultimate goal of sample selection is to obtain samples of participants and non-participants representative of the population of interest. How exactly that is achieved depends on how the sample selection step relates to the randomization step. We can think of two limiting cases:

1. **Randomization precedes sample selection:** In this case, selection of samples of participants and non-participants occurs *after* random assignment. Randomization has already

[52] If not, inference about program impact for the population of interest from a sample might be considerably complicated. It is difficult to offer general recommendations about how to proceed with sample selection in that instance since there are so many ways that the two could conceivably not be the same.

[53] The case where the participation assignment mechanism was not deliberate but participation was nonetheless fully randomized is, from the standpoint of causal identification of program impact, effectively the same as an RCT. Such a case would probably fall under the category of "natural experiments," though this terms is more often applied when there was some source of random variation in participation that did not result in full randomization of participation.

[54] Where by "post"-participation we mean simply any time period after participation commences.

insured that participants and non-participants are the same on average in terms of their pre-participation characteristics. The goal of sample selection is to obtain samples representative of the participants and non-participants in the RCT. By selecting representative samples the main benefit of the randomization of program assignment (that participants and non-participants differ on average only by the experience of the program) is preserved in the samples so obtained. This circumstance usually arises when the participants and non-participants are far more numerous than impact estimation requires. For instance, one might think of the case where the residents of a state are randomly assigned to receive or not receive a program. Actual impact evaluation might require a far smaller sample of individuals participating and not participating.[55]

2. **Sample selection precedes randomization:** In this instance, a representative sample from the population of interest is obtained. Then, the members of that sample are randomly assigned to receive or not receive the program. This is often the practice when the randomized controlled trial is in the vein of a limited trial or pilot implementation of the program. When sample selection comes first, the sample is (if selection was done in an appropriately random fashion) representative of the population of interest. Random assignment to participant and non-participant status then insures that participants and non-participants are subsamples also representative of that population.[56]

In our work, we have encountered both cases. There are also cases that in some respect or another fall between these two possibilities.

Regardless of which of these two cases one is confronted with, the goal of sample size estimation is, for all the complicated math, pretty simple: to insure samples of participants and non-participants are of sufficient size to meet statistically the impact evaluation goals. And we have seen what these are — to test the hypotheses

$$H_0 : \overline{I} = 0$$

$$H_a : \overline{I} \geq d \text{ or } \overline{I} \leq -d$$

with power $1 - \beta$ and significance α. Whatever the relationship between randomization and sample selection for the evaluation survey, the focus must be kept on this sampling goal.

[55] One might ask how it is that one can be sure the RCT is conducted for the population of interest. There are two possibilities. First, it might be that randomization occurred with respect to every member of the population of interest. In that case, everyone in the population ends up either a participant or non-participant. Second, it might be that some representative sample was chosen from the population of interest which was then subjected to randomization (that is, participated in the RCT). The point is that selection of the sample in this first stage was not guided by the sampling goals that will drive selection after randomization. The purpose of the first stage randomization may have been simply to reduce the scope of the RCT in terms of the number of participants and non-participants it involved. Whatever the case may be, this case is motivated by the circumstance where randomization has already occurred and the set of individuals subject to randomization per the RCT is determined before sample selection.

[56] One question is how randomization should occur. There are many ways of doing so, but in all it is preferable if it is done in such a way that the numbers of participants and non-participants is guaranteed (to the extent that it can be by investigators). This preserves from a technical standpoint the unbiasedness of estimates of average outcomes for participants and non-participants.

RCTs are often conducted on a cluster basis. By this we can (and in practice typically do) mean two things: that randomization is done at some kind of spatial level, and that sampling (at least in the initial stages of a multistage selection process) is as well. Randomization at the level of some spatial unit can be motivated by a couple of considerations. For instance, it could be that the intervention is delivered at some spatial level (as in the placement of a clinic to treat those in some reference catchment area, in which case those in that catchment area would be regarded as "treated" or "participating").

Cluster-based designs for surveys of participants and non-participants for the purpose of estimating impact are typically attractive for their cost savings. For instance, if one wishes to obtain a sample with 10,000 individuals, it is typically cheaper to interview 25 individuals in each of 400 clusters than to actually independently select and interview 10,000 individuals scattered in 10,000 locations.[57] However, there is a cost to this. The sample of 10,000 individuals in 400 clusters typically contains less true independent variation than the sample of 10,000 individuals selected individually.

Implementation of a cluster-based sample selection design is typically in the context of a multistage selection process. Often, the first stage of selection involves some kind of selection of spatial units, followed sometimes by further selection of smaller spatial units within the larger units selected at the preceding stage of selection. Eventually, units of analysis (including individuals, households, etc.) are selected. In the case of RCTs with spatial randomization, this suggests the additional attraction to cluster-based sampling that in many instances the members of the spatial sampling units at the final stage of selection for the spatial sampling units might tend to be all of a kind (participants or non-participants) depending on the random assignment per the RCT of the area within which the spatial unit falls. This can often make it easier to insure particular sample sizes of participants and non-participants.

As mentioned, the spatial units of randomization and sampling do not need to be the same. For instance, one of the authors of this manuscript is presently involved in a study wherein participation in a program is allocated across relatively high-level administrative units of a region of a country, but actual sampling to obtain samples of participants and non-participants will involve a first stage of selection whereby samples of villages, a much lower level administrative unit, are selected from the administrative units (randomly) assigned to participant or non-participant status.

That said, when one refers to a cluster-randomized design, (s)he usually means that randomization of program assignment occurs at a lower level. For instance, one might select some clusters from the universe of clusters containing a population of interest to participate in the randomized controlled trial and then randomly assign these selected clusters to either participate or not participate in the program. This was more or less the foundation of, for instance, the famous Progresa randomized controlled experiment evaluation. The actual surveys to obtain samples to evaluate the program drew households from all clusters involved in the Progresa evaluation.

For the purpose of sample size estimation, what is really important is how the samples of

[57] To be fair, by random chance there would probably be instances where by random chance selected individuals lived close to one another, but the basic point is still valid.

participants and non-participants are selected. When there will be no "design" (beyond simple independent selection of units of analysis — in other words, no cluster-based selection — with no unequal probabilities of selection and no stratification) sample size is estimated by the straightforward formulas from the last section whereby sample size for the continuous case is per the formula

$$n = 2 \cdot \left(\left[\frac{\sigma_Y \cdot \left(z_{1-\alpha/2} + z_{1-\beta} \right)}{d} \right]^2 \right)$$

whereas sample size for the case where the outcome is a rate or proportion is

$$n = \left[\frac{z_{1-\alpha/2} \cdot \sqrt{2 \cdot \overline{\overline{Y}} \cdot \left(1 - \overline{\overline{Y}}\right)} + z_{1-\beta} \cdot \sqrt{\overline{Y}_1 \cdot \left(1 - \overline{Y}_1\right) + \overline{Y}_2 \cdot \left(1 - \overline{Y}_2\right)}}{d} \right]^2$$

In the last section we developed extensions of these formulas for cases where the sample sizes for participants and non-participants are not the same and where the outcome might be correlated between participants and non-participants.

Cluster-based sampling means generally, however, that there will be a design. The typical way to handle this in cluster-based randomized controlled trials is to estimate the sample size as

$$n^{deff} = deff \cdot n$$

where n is the sample size per the formulas of the last paragraph and $deff$ is the design effect generated by the way that the sample is selected. A very common approach to sample size estimation for randomized controlled trials is to estimate $deff$ in this way

$$deff = 1 + (M - 1) \cdot \rho_{ICC}$$

where M is the sample size per cluster and ρ_{ICC} is the intracluster correlation (Kish, 1965).

This design effect estimate is motivated by the idea that the only design feature that departs from the classical selection assumption is the non-independence of observations as a result of cluster-based selection. If the authors experience is any guide, other design features will likely be a factor and hence should inform the design effect estimate. For instance, in our experience multistage cluster-based sampling rarely, in fact, leads to a self-weighting sample. A prior for the design effect based on more than just potential clustering would be appropriate in cases of uneven probabilities of selection and stratification. The ideal (at least in the authors' experience) is when a prior for the design effect can be formed with estimates of the design effect from an (ideally recent) existing sample from the population of interest that was selected in a fashion similar to that contemplated for the survey to support estimation of impact in light of the looming RCT.

All of this being said, the discussion to this point would seem to suggest that, on the whole, sample size estimation for RCTs (clustered or not) is quite simple. There are, however, two wrinkles to the straightforward application of the sample size estimators of the last section to samples from populations participating in randomized controlled trials.

First, with samples from RCTs, impact is often estimated in more sophisticated ways. There might be estimation possibilities (usually some form of multiple regression) that might offer more efficient estimates of impact even when random assignment to participation status is obeyed and the sample of participants and non-participants is thus experimental. Moreover, RCTs don't always work out (in the sense of delivering truly experimental samples). The usual reason for this is that the members of the population of interest simply do not fully cooperate with their random assignment. Sometimes the program implementers do not cooperate with the randomized assignments, while in many other cases it is the participants in the RCT who do not cooperate with their random assignment. Once this happens, the main advantage of the randomized control design (randomly assign to program participation status insuring that participants and non-participants differ only by the experience of program participation, and thus that any differences in outcomes between the two subpopulations reflects the impact of program participation) is lost and we are left with essentially non-experimental samples for which we must appeal to quasi-experimental strategies for estimation of program impact. Sample size estimation for these quasi-experimental estimators is considered in subsequent sections.

Our general advice at this point is that when contemplating the design of an RCT (including the sample selection design to obtain samples of participants and non-participants) to estimate sample size is to perform sample size estimation not only for the basic comparison of averages approach developed in the last section, but also for whatever quasi-experimental design(s) the investigators envision using in the event of less than complete cooperation with randomized assignment. Conservative sample size estimation would then dictate that the largest of the sample sizes recommended by these various sample size estimation exercises should be pursued.

Second, some randomized controlled trial designs are more complicated than simply randomly assigning some to participate and some to not participate. Once one starts thinking about this, there could be many ways of potentially randomizing for the purpose of pursuing estimates of different population program impact parameters. And indeed there are many variations that have been applied to estimation of causal relationships. We briefly mention one particular extension of the basic RCT design that has proven increasingly popular: the step wedge design.

The original motivation for the step wedge design was likely less the pursuit of some particular population program impact parameter or statistical advantage than an ethical consideration. Specifically, regardless of their statistical benefits for the purpose of identifying program impact, RCTs have often been a tough sell from an ethical standpoint. By design, a basic RCT (whereby some are simply assigned to participate and some are not) involves the prospect of denying some the opportunity to participate in a potentially beneficial program (one would hope that the program would not be pursued, even on a trial basis for an RCT, unless there was some reasonable hope that it provided some kind of benefit[58]).

The basic step wedge design involves sequential assignment to program participation. Essen-

[58] Of course, program participation could admittedly involve some risks as well.

tially, the population of interest (or some sample from it) is randomly assigned to groups. Group assignment determines the order in which participation occurs. In the typical (or maybe idealized) step wedge design, no one initially participates in the program. Groups sequentially take up participation until at some point everyone is participating. This basic mechanism is illustrated in Table 3.3. In the hypothetical example in Table 3.3, the population (or some sample from it) is randomly assigned to one of seven groups. Beginning from an initial time period (time period 1) where no one is participating, participation is gradually extended to the seven groups sequentially until everyone participates. The design is called the "step wedge" because cumulative enrollment over time looks rather like a cross section of a series of steps (as in Figure 3.12, where the red blocks each represent a group participating in the program at the indicated time period while the white blocks represent a group not yet participating as of the indicated time period).

Table 3.3. A step wedge design

(YES=currently participating; NO=not currently participating)

	Time							
Group	1	2	3	4	5	6	7	8
1	NO	YES	YES	YES	YES	YES	YES	YES
2	NO	NO	YES	YES	YES	YES	YES	YES
3	NO	NO	NO	YES	YES	YES	YES	YES
4	NO	NO	NO	NO	YES	YES	YES	YES
5	NO	NO	NO	NO	NO	YES	YES	YES
6	NO	NO	NO	NO	NO	NO	YES	YES
7	NO	NO	NO	NO	NO	NO	NO	YES

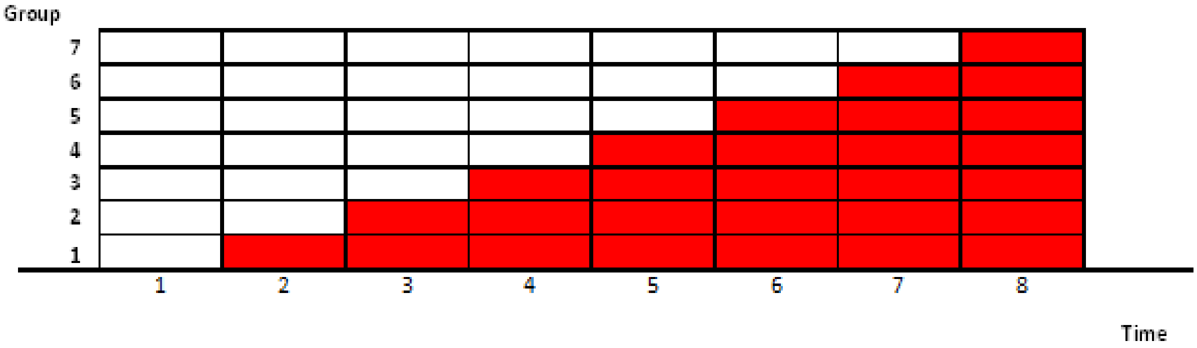


Figure 3.12. Cumulative enrollment of groups under the step wedge design

The ethical attraction of the step wedge design should now be apparent: ultimately, no one is denied access to a potentially beneficial program. That said, the step wedge design could potentially open up the opportunity to consider other population impact parameters, such as parameters that somehow reflect dose response possibilities. In the case of the step wedge design, the "dose" would be the time of exposure to or participation in the program. If there were a particular time of exposure question one might want to apply the basic comparison of averages between groups sample size estimation strategies of the last section to insure sufficient sample size. For instance, the impact of one period of participation against seven could, in the context of Table 3.3, involve comparison of average outcomes between those in group 1 and those in group 7. If estimation of the dose response as reflected in the difference in average outcomes between these two groups was important, it would be a good idea to insure via the methods of the last section that the sample sizes in the two groups were sufficient to detect the desired dose response effect with sufficient power and significance.

Given the mounting popularity of step wedge designs in the past decade or so, it is perhaps unsurprising that a literature on sample size estimation for them has gathered steam. As with sample size estimation for simpler cluster RCTs, the general scheme for sample size estimation tends to involve estimation of the sample size under simple selection multiplied by the design effect:

$$n^{deff} = deff \cdot n$$

The focus of the sample size estimation literature for step wedge designs is then forming an appropriate prior for $deff$.

Given that this seems to be a fast-moving literature in play as we write this, we discuss one very recent contribution to it (Hemming & Taljaard, 2016). Heming and Taljaard attempt to provide a “unified” approach to sample size estimation. Along with the step wedge design, they consider two other cases: basic cluster randomized designs and a pre-/post-, experimental-/control- cluster design. To fix ideas, these are illustrated visually in the Figures 3.13 and 3.14 (we refer to the pre-/post-, experimental-/control- cluster design as the “before-after cluster RCT” in Figure 3.14) to illustrate them in the same terms as step wedge designs in Figure 3.12. In Figure 3.13 there is only one time period of observation and 4 of the 7 groups are randomly assigned to participate. In Figure 3.14 there are two time periods: one before program implementation and one after, when 4 of the 7 groups are randomly assigned to participate.

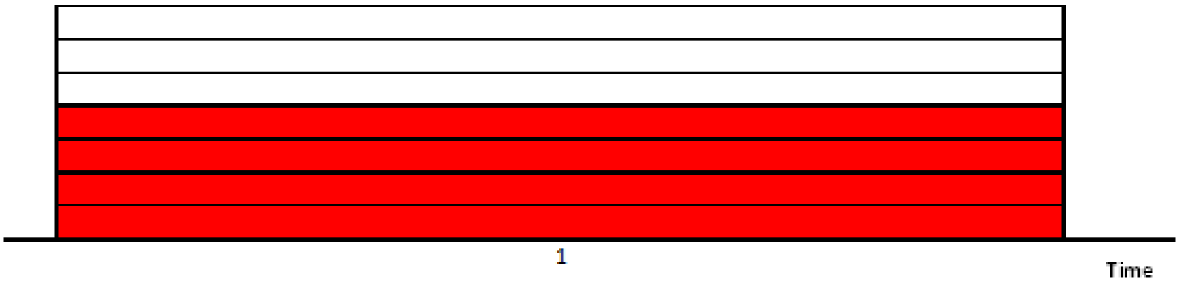


Figure 3.13. A cluster RCT

They begin with a familiar finding: the sample size under the cluster randomized control design is

$$n^{deff} = deff \cdot n$$

where n is the sample size per the formulas of the last paragraph and $deff$ is the design effect generated by the way that the sample is selected and $deff$ is

$$deff = 1 + (M - 1) \cdot \rho_{ICC}$$

The notional backdrop for this is that the clusters have been selected and randomization occurs subsequent to cluster selection. Our concern about this approach remains: its implicit agnosticism regarding what is happening in the cluster (such as within-cluster random selection potentially giving rise to samples that are not self-weighted) means that this might not be a reasonable estimator of the design effect for every application. This formula essentially assumes that the sample size per cluster M is predetermined, so that the number of clusters c is then

$$c = \frac{n^{deff}}{M}$$

The predetermined nature of the cluster size M highlights our concern about a general rule of using $1 + (M - 1) \cdot \rho_{ICC}$ as our estimator of the the design effect: if the clusters vary in size there will be unequal probabilities of selection. Heming and Taljaard (2016) note that this design is not advised when

$$c < n \cdot \rho_{ICC}$$

However, having a number of clusters c of

$$c = n \cdot \rho_{ICC}$$

is not guaranteed to be sufficient. Various rules of thumb regarding the appropriate number of clusters have been advanced, though we cannot speak to the general advisability of them.

If one wishes instead to make the number of clusters c the predetermined variable then the

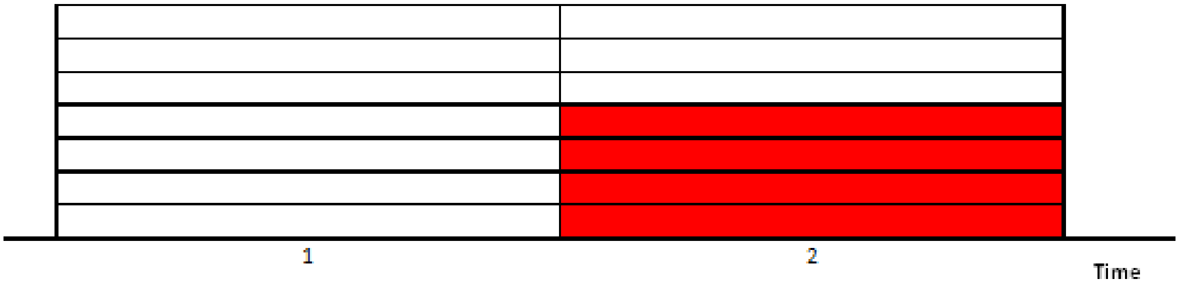


Figure 3.14. A before-after cluster RCT

cluster size is[59]

$$M = \frac{n \cdot (1 - \rho_{ICC})}{c - n \cdot \rho_{ICC}}$$

This approach still involves the concerns highlighted in the last paragraph.

Moving onto the before-after cluster RCT, Heming and Taljaard recommend the design effect estimator

$$deff = 2 \cdot \left(1 + \left(\frac{M}{2} - 1\right) \cdot \rho_{ICC}\right) \cdot \left(1 - r^2\right)$$

where r is the correlation between cluster means over the two time periods:

$$r = \frac{\frac{M}{2} \cdot \rho_{ICC}}{1 + \left(\frac{M}{2} - 1\right) \cdot \rho_{ICC}}$$

r is the same as ρ from the basic comparison of correlated sample means formula of the last section.[60] This formula assumes the cluster sample size M is predetermined, in which case solving for the number of clusters c is straightforward. The process for obtaining M when c is the predetermined parameter is a bit more complex, but not of as much interest (vary rarely is it c that is predetermined; see Heming and Taljaard [2016] for further details on this case).

Finally, we arrive at the step wedge design, the design effect for which is

$$deff = (t + 1) \cdot \frac{1 + \rho_{ICC}\,(t \cdot M + M - 1)}{1 + \rho_{ICC} \cdot \left(\frac{t \cdot M}{2} + M - 1\right)} \cdot \frac{3 \cdot (1 - \rho_{ICC})}{2 \cdot \left(t - \frac{1}{t}\right)}$$

where t is the number of steps. This design effect estimator also sidesteps the question of unequal selection probabilities from within cluster selection.

There are of course many different variations on the RCT. Consider, for instance, the randomized block design. Essentially, this involves sorting those who will participate in the randomized controlled trial by their type (the resulting groups of participants in the randomized controlled trial are "blocks"). Randomization into program participant and non-participant groups then proceeds in each block.

We did not discuss sample size estimation for the block randomized design for reasons that illustrate why we go no further in our discussion of sample size estimation for successful randomized controlled trials. First, although some options for sample size estimation have been suggested for randomized block designs, we are unaware of any conventional standard of sample size estimation (a

[59] To see this, note that

$$n^{deff} = deff \cdot n$$

is

$$c \cdot M = (1 + (M - 1) \cdot \rho_{ICC}) \cdot n$$

Then simply solve for M.

[60]

$$n = 2 \cdot (1 - \rho) \left(\left[\frac{\sigma \cdot \left(z_{1-\alpha/2} + z_{1-\beta}\right)}{d} \right]^2 \right)$$

"gold standard," if you like) as widely accepted as the simple comparison of means formulas of the preceding section. Second, the sample size estimators we know of are not simple or straightforward in terms of their foundations: the discussion and derivation of some of the alternatives we have seen would not be a short or simple affair. For instance, one of the frequently-mentioned approaches would require a thorough review of two-way ANOVA before we could even get to formula derivation. Third, in our (admittedly highly subjective) assessment, the sample size estimators we have seen are not particularly persuasive. They require a large number of assumptions, both in derivation and for actual sample size estimation in practice. Finally, and perhaps most importantly, there are so many variations on the randomized block design (with missing blocks, imbalanced blocks, crossover, etc.) and extensions of the basic block randomized design (e.g., Latin Squares) that there would really be no end to the discussion.

What essentially all of these variations on and extensions of the randomized block design have in common is that they claim some efficiency advantage over straightforward randomization. In some sense, the sample size estimation formulas available ask users to make strong assumptions about the nature of that efficiency savings (making the results of the sample size estimation less confidence-inspiring). A relatively conservative approach is thus simply to use conventional, straightforward sample size estimation formulas for simple comparison of means. This will produce comparatively large sample size estimates, which provide good insurance against the potential failure of the assumptions of the more elaborate approaches.

As a last note on RCTs (at least for now) we touch on a particularly troublesome issue: what if the participants and non-participants are not representative of the population of interest? As samplers, the most likely way that we would be presented with this is by being presented with a group of (already selected) participants for the RCT who are somehow not representative of the population that the impact evaluation intends to address. Even if this is the case, it is still important that selection of samples of those randomly assigned to participate and not participate yield representative samples of participants and non-participants. Those samples will then still be representative of whatever population the participants in the randomized controlled trial do represent. It is possible that measures such as post-stratification adjustments could then move estimates of impact closer to being representative of the population of interest in terms of observed characteristics, though this is an issue to which we will return in the next chapter.

While our focus in this section has been RCTs and hence experimental samples, the impact estimation strategies and sample size estimation approaches could be applied to non-experimental samples as well. Sometimes this is in the wake of an RCT with less than full compliance. More generally, simple comparison of outcomes between observational (i.e., drawn from real world circumstances) samples of participants and non-participants, with no presumption of any attempt at randomization of participation, is probably the most widely-used strategy to date for estimating program impact with samples of participants and non-participants.

The challenge in the case of non-experimental data lies with the interpretation of the estimates of impact so obtained. In the absence of randomization of participation, observed participants and non-participants likely differ on average by all of the factors that guided the non-random determination of participation. This makes it very difficult to ascribe observed differences in average outcomes between samples from the two groups to the experience of program participation. We cannot say what part of any estimated difference is due to program participation and what part is due to the fact that different types of individuals ended up among the participants and non-participants.

Having said this, there is no reason the impact estimation methods, and sample size estimation strategies, discussed in this section and the last section, cannot be applied in the circumstance of non-experimental data. For instance, none of the "comparison of mean outcomes" sample size estimators described in the last section rely intrinsically on the idea that membership in the two groups/populations is randomly determined (and indeed, it could be incredibly "non-random," as in comparing outcomes between samples from populations of the young and old, men and women, residents of different regions, etc.).

A special case to be aware of is the simple "pre-/post-" design. This is a type of non-experimental design whereby no one participates at some baseline observation point and everyone (or at least everyone observed) does at some follow-up time period. This is a fundamentally non-experimental design because it cannot control for any factors beyond participation that influence the outcome and may have evolved from baseline to follow-up. For sample size estimation purposes, what is important to recognize is that in any case where the sample design has any kind of longitudinal dimension, there is likely to be some correlation of outcomes between baseline and follow-up for which some accounting must be made. For instance, in the continuous outcome case the appropriate sample size would be given by

$$n = 2 \cdot (1-\rho)\left(\left[\frac{\sigma_Y \cdot \left(z_{1-\alpha/2} + z_{1-\beta}\right)}{d}\right]^2\right)$$

whereby we capture the intertemporal correlation of outcomes via ρ.

3.3 Selection on Observables

We now turn to the subject of sample size estimation for selection on observables methods for estimating program impact. The selection on observables approach to estimation of causal relationships (such as that between program participation and some outcome of interest) spans a wide range of actual estimation strategies. Broadly, it involves two major traditions, regression and matching (and derivatives of matching, such as inverse probability weighting approaches). Within these two traditions are a vast number of possible specific estimators (think just of how many approaches to propensity score matching have been proposed).

What these many different specific estimation strategies have in common is their core identifying assumption. Specifically, they assume that we observe and control for all of the factors that guide potential outcomes and the participation decision.

Consider the potential outcomes $\{Y^1, Y^0\}$. The identifying assumption (i.e., the assumption that allows estimates of program impact to be interpreted as the causal impact of the program on outcomes) of randomized controlled trials is that program participation is randomly determined. If this is the case, program participation P should be independent[61] of the potential outcomes $\{Y^1, Y^0\}$. Then, we have

$$E\left(Y^1|P=1\right) = E\left(Y^1|P=0\right) = E\left(Y^1\right)$$

and

$$E\left(Y^0|P=0\right) = E\left(Y^0|P=1\right) = E\left(Y^0\right)$$

Because these are true (under the assumption of randomization) we can estimate the population average outcome under participation with a sample of participants (because randomization insures that they represent the entire population). Similarly, we can estimate the population average outcome in the absence of participation with a sample of non-participants. In principle, we can then estimate program impact as a simple difference of these estimates.

Randomization, which decoupled program participation from potential outcomes, was the key to this straightforward estimation possibility. Suppose instead that program participation was not randomly determined, but instead the product of purposeful decisions. This could reflect either the decision-making of individuals or the allocation decisions of program planners and administrators (or both). Whatever the case, this opens the door to the possibility (in fact likelihood) that program participation is influenced by and hence associated with some characteristics of the individual participants (these can be the immediate personal characteristics of those individuals, of their households, of their larger community context, etc.). However, if any of these characteristics influence potential outcomes then participation is associated with those potential outcomes.

But then, in general

$$E\left(Y^1|P=1\right) \neq E\left(Y^1|P=0\right) \neq E\left(Y^1\right)$$

and

$$E\left(Y^0|P=0\right) \neq E\left(Y^0|P=1\right) \neq E\left(Y^0\right)$$

The simple estimation strategy of randomized controlled trials would then break down. For instance, the experiences of a sample of participants no longer estimate the population average outcome under participation for either non-participants or the combined population of participants and non-participants.

Selection on observables impact evaluation estimation approaches assume that the factors that influence the program participation decision and potential outcomes are observed. These methods then seek to identify program impact by controlling for these factors as determinants of potential outcomes so that the remaining variation in potential outcomes associated with program participation is independent of those potential outcomes (in other words, is analogous to the variation on which randomized controlled trials rely).

[61] Actually, in general, randomized controlled trials hinge on the somewhat weaker assumption of mean independence of program participation and potential outcomes.

To fix ideas, suppose that we are concerned with one characteristic X that influences both program participation P and potential outcomes $\{Y^1, Y^0\}$. The identifying assumption of selection on observables models can be stated as

$$E\left(Y^1|X,P\right) = E\left(Y^1|X\right)$$

and

$$E\left(Y^0|X,P\right) = E\left(Y^0|X\right)$$

In other words, after conditioning on X program participation P reveals no information about the potential outcomes. The remaining average differences in the outcome

$$Y = P \cdot Y^1 + (1 - P) \cdot Y^0$$

between participants and non-participants can then be ascribed to the experience of program participation (as opposed to some intrinsic difference in their average potential outcomes).

This is a pretty strong departing assumption. It essentially precludes the possibility of an *unobservable* factor or characteristics that influences both program participation and potential outcomes: if we cannot observe this factor then we cannot control for it per the methods of the selection of observables estimation strategies. This means that the remaining variation in P would not necessarily be independent of potential outcomes, hence differences in observed outcomes between participants and non-participants (and representative samples from them) would not reflect simply the experience of program participation (i.e., program impact) but also differences in the distribution of those unobserved factors between participants and non-participants. For instance, if motivation influences both participation and potential outcomes, then differences in outcomes between participants and non-participants represent program impact and the differences in motivation levels of participants and non-participants.

In general, assuming that all such potential "confounders" are observed and hence can be controlled for is a strong one. Nonetheless, the selection on observables approaches are extremely popular program impact estimation methods. Indeed, we would hazard to guess that, taking the body of work to evaluate programs as a whole, selection on observables methods are probably second in popularity only to evaluations that identify impact by simple comparison of average outcomes between participants and non-participants.

3.3.1 Regression

3.3.1.1 Linear Regression

We begin with linear regression, from which the workhorse sample size estimation method for selection on observables methods emerges. Suppose that we have a sample of size n from some population of interest for program impact estimation purposes (in other words, we wish to know program impact for that population). For each individual in that sample we observe

$$\{Y_i, P_i, X_i\}$$

where Y_i is the observed outcome of interest for individual i for $i = 1, ..., n$, P_i is program participation for individual i, and X_i is an observed characteristic of individual i. X_i could be thought of as either a single characteristic of individual i or a vector of them. When we say characteristics of individual i, we mean it in the broadest sense: the characteristics could be attributes of the individual, his/her household or his/her larger community or environment. The linear regression approach then involves regression of Y_i on P_i and X_i using the each of the n individuals in the sample. The regression model implicitly behind this is

$$Y_i = \beta_0 + \beta_1 \cdot P_i + \beta_2 \cdot X_i + \varepsilon_i$$

The estimated coefficient on program participation P, $\hat{\beta}_1$, is then typically interpreted as the estimated marginal effect of program participation on the outcome Y.

It is worth thinking about how this regression model might arise. We begin with a simple behavioral model motivated by the potential outcomes framework. Specifically, we have the potential outcome equations

$$Y_i^0 = \beta_0 + \beta_2 \cdot X_i + \varepsilon_i$$

and

$$Y_i^1 = \beta_0 + \beta_1 + \beta_2 \cdot X_i + \varepsilon_i$$

where X_i is defined above, Y_i^1 is the outcome individual i experiences if they participate, Y_i^0 is the same if they do not participate, and ε is an unobserved characteristic of the individual (in other words, $\{X, \varepsilon\}$ take on different values for different individuals but only the variation in X is actually observed across individuals). Program impact for individual i is then

$$Y_i^1 - Y_i^0 = \beta_1$$

This is thus a homogeneous treatment effect framework (we will return briefly to that below).

We also specify the cost of participation as

$$C_i = \gamma_0 + \gamma_1 \cdot X_i + \varepsilon_i^C$$

Thus, the cost of participation now depends on the observed variable as well as an unobserved characteristic represented by ε_i^C. We adopt the earlier condition for determining participation status. Specifically, the individual participates (i.e., $P_i = 1$) if

$$Y_i^1 - Y_i^0 - C_i > 0$$

or, inserting the functions we have proposed,

$$\beta_0 + \beta_1 + \beta_2 \cdot X_i + \varepsilon_i - \beta_0 - \beta_2 \cdot X_i - \varepsilon_i - \gamma_0 - \gamma_1 \cdot X_i - \varepsilon_i^C \geq 0$$

or, removing terms,

$$\beta_1 - \gamma_0 - \gamma_1 \cdot X_i - \varepsilon_i^C \geq 0$$

Rearranging, we have

$$-\gamma_1 \cdot X_i - \varepsilon_i^C \geq -\beta_1 + \gamma_0$$

Whether this inequality holds clearly depends on the values of X_i and ε_i^C. Thus, this is a framework where program impact varies across individuals, and certain observed and unobserved types of individuals are more common among participants than non-participants.

There are two fundamental questions one could ask at this point. First, one could ask why anyone would ever choose to participate if

$$Y_i^1 - Y_i^0 = \beta_1 = 0$$

In other words, why would we observe anyone participating if the true impact from doing so was zero?[62] At first this might seem to be an artifact of assuming a constant program impact β_1, which might be zero. However, even in a world of heterogeneous treatment effects (i.e., impact varying across individuals) there is still the hypothetical possibility of someone participating even if the impact of the program on them is zero. A first response might be to assume non-negative costs (i.e., impose the restriction $C_i > 0$). The problem with doing this is that then *no one* would participate if $\beta_1 = 0$. Even allowing $C_i = 0$ for some and then assuming that such individuals would participate (i.e., assuming $C_i \geq 0$ and weakening the participation condition to $Y_i^1 - Y_i^0 - C_i \geq 0$) seems rather like mathematical tinkering rather than a persuasive behavioral explanation.

A more behaviorally persuasive answer involves a more holistic approach to costs. All sorts of factors probably enter into an individual's reckoning of the costs: the monetary cost, the time cost in terms of foregone labor, any psychic costs in terms of stress, lost time with family and friends, missing reruns of *Frasier*, etc. But there also factors that might be negative. These might include, for instance, the social benefit or satisfaction of participation. These could be negative, leading to a negative cost for a given individual. Or perhaps the program improves some other channel of human welfare. More generally, program participation could convey benefit in terms of outcomes other than Y that could lead to negative costs C_i. In other words, some individuals would rather enroll even if their payoff in terms of the outcome of interest, $Y_i^1 - Y_i^0$, was zero because they gain in other ways.

[62] This is distinct from the situation where we fail to reject the null hypothesis

$$H_0 : \beta_1 = 0$$

based on testing using estimates from a sample. In that case observing that some participated is not so concerning since, in empirical practice, we likely had low power to reject alternative hypotheses involving very small values for β_1.

Another question is how to square this atomistic, individual agency-rooted approach to participation with the reality that participation is determined by program officers, government officials or other social planners operating above the individual participant level. An easy way to think about this is that the program or government personnel are operating as an agent of the individual. Hence they attempt to solve the individual's problem for them. In cases where program allocation decisions are made at the community level, this typically involves solving the participation problem for the typical person in each community.

Observed Y_i is

$$Y_i = P_i \cdot Y_i^1 + (1 - P_i)\, Y_i^0$$

$$= P_i \cdot (\beta_0 + \beta_1 + \beta_2 \cdot X_i + \varepsilon_i) + (1 - P_i) \cdot (\beta_0 + \beta_2 \cdot X_i + \varepsilon_i)$$

$$= \beta_0 + \beta_1 \cdot P_i + \beta_2 \cdot X_i + \varepsilon_i$$

We thus have a linear model of the observed outcome Y_i in terms of all the observed factors (P_i and X_i) and a stochastic element (ε_i).[63]

Assuming that we can estimate the parameters β_0, β_1 and β_2 (but, as we will shortly see, particularly β_1) in an unbiased fashion, we obtain an unbiased estimate of program impact. To see this, let's begin with what program impact is in the context of the behavioral model. In other words, lets see what the potential outcomes framework tells us about true program impact. We have

$$\frac{\sum_{i=1}^{n} Y_i^1 - Y_i^0}{n} = \frac{\sum_{i=1}^{n} \beta_0 + \beta_1 + \beta_2 \cdot X_i + \varepsilon_i - (\beta_0 + \beta_2 \cdot X_i + \varepsilon_i)}{n} = \frac{\sum_{i=1}^{n} \beta_1}{n}$$

$$= \frac{n \cdot \beta_1}{n} = \beta_1$$

and

$$\frac{\sum_{j=1}^{N} Y_j^1 - Y_j^0}{N} = \frac{\sum_{j=1}^{N} \beta_0 + \beta_1 + \beta_2 \cdot X_j + \varepsilon_j - (\beta_0 + \beta_2 \cdot X_j + \varepsilon_j)}{N} = \frac{\sum_{j=1}^{N} \beta_1}{N}$$

$$= \frac{N \cdot \beta_1}{N} = \beta_1$$

where N is the population size (and N could be ∞). In other words, true program impact in both the sample and the population is β_1.

One might ask why, if this is the case, we estimate impact from the sample (as opposed to compute it exactly). There are two sources of uncertainty about impact at the population level given the information in the sample. First, and most, important, we cannot observe $\{Y_i^1, Y_i^0\}$ for our sample. Instead, for each individual i we can observe at most one of the two: Y_i^1 for participants ($P_i = 1$) and Y_i^0 for non-participants ($P_i = 0$). So we could never mechanically calculate impact from the potential outcomes framework in the fashion of the last paragraph. Second, and essentially because of this, we must focus on the observed outcome Y_i for which we cannot remove the role of

[63] Some have considered the possibility that P_i and X_i are themselves random variables, as opposed to fixed quantities. In some sense, this is a logical extension since P_i is random in our model (for example, it depends in part of ε_i^C. We will discuss this briefly later in this subsection.

the random determinant ε_i of the observed outcome

$$Y_i = \beta_0 + \beta_1 \cdot P_i + \beta_2 \cdot X_i + \varepsilon_i$$

This uncertainty, which is intrinsically an artifact (at least under this model) of the partial observability of potential outcomes, places us in the realm of estimation.

Nonetheless, the coefficient on program participation in the regression model, β_1, is program impact. We thus want that to be the estimand we are pursuing in our estimation of the parameters of this regression model. Therefore, the objective is to form estimates $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ that would be unbiased:

$$E\left(\hat{\beta}_0\right) = \beta_0$$

$$E\left(\hat{\beta}_1\right) = \beta_1$$

and

$$E\left(\hat{\beta}_2\right) = \beta_2$$

where the second condition is of particular importance since estimating program impact is our focus.

Our focus in this subsection is estimation of β_0, β_1 and β_2 (but, as we have emphasized, β_1 in particular) via linear regression such as ordinary least squares. Unbiased estimation of these parameters requires, above all, the independence of ε_i and ε_i^C (see Lance, Guilkey, et al., 2014 for a discussion of this, as well as the consequences of weakening this assumption). We proceed in this subsection (and indeed in this entire section) as if that assumption is reasonable. If it is not, then selection on observables methods are not, *per se*, reasonable.

In terms of interpretation via the full set of linear regression estimates ($\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$,), program impact is typically approached as the marginal effect of the program. In the face of program participation defined in a binary fashion as participate/don't participate (which is typical in program impact estimation), the marginal effect of program participation is defined by

$$E\left(Y_i|P_i = 1, X_i\right) - E\left(Y_i|P_i = 0, X_i\right)$$

Here we define the marginal effect at the individual level. Assuming unbiased estimation of β_0, β_1 and β_2 via linear regression, we have

$$E\left(Y_i|P_i = 1, X_i\right) - E\left(Y_i|P_i = 0, X_i\right) = E\left(\hat{\beta}_0 + \hat{\beta}_1 \cdot 1 + \hat{\beta}_2 \cdot X_i + \varepsilon_i\right) - E\left(\hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot X_i + \varepsilon_i\right)$$

$$= E\left(\hat{\beta}_1\right) = \beta_1$$

assuming unbiased estimation. Thus, assuming unbiased estimation of the regression parameters β_0, β_1 and β_2, $\hat{\beta}_1$ is an estimate of program impact.

It is fairly easy to extend this to consider heterogeneous treatment effects. For instance, we could assume that

$$Y_i^1 = \beta_0 + \beta_1 + \beta_2 \cdot X_i + \varepsilon_i^{Y1} + \varepsilon_i$$

where ε_i^{Y1} is, effectively, a "random coefficient" factor for program participation. True average program impact would then be

$$E\left(\beta_2 + \varepsilon_i^{Y1}\right) = \beta_2$$

In this instance, simple regression of Y_i on P_i and X_i is not guaranteed to yield an unbiased estimate of β_1 (indeed, it is actually hard to see how it would under a behaviorally reasonable model; see Lance, Guilkey, et al. [2014]). For now, however, we ignore the possibility of heterogeneous treatment effects.[64]

Linear regression is a selection on observables method because it allows for the possibility that an observable X_i is correlated with both the outcome Y_i and program participation P_i. If we did not control for X_i, it would serve as a confounder that would make it impossible to ascribe differences in the average of Y_i between participants ($P_i = 1$) and non-participants ($P_i = 0$) to the impact of the program. In that event the difference in average outcomes would reflect not only program impact but also differences in X between the two groups. Linear regression solves this problem by simultaneously controlling for P_i and X_i in working out the contribution of each to shaping Y_i. This also explains conceptually why the independence of ε and ε^C is so important — it precludes the possibility of *unobservables* that determine the outcome Y and program participation P.

To summarize, if ε and ε^C are independent, then

$$E\left(\hat{\beta}_1\right) = \beta_1$$

In other words, the estimand of the coefficient on program impact is the population average program effect from the potential outcome framework. Furthermore, $\hat{\beta}_1$ is also the most natural measure of program impact under the regression model (i.e., the marginal effect of program participation). Sample size estimation for linear regression as an impact estimator would then naturally center around sampling goals in terms of what we want to learn about β_1 from the estimate $\hat{\beta}_1$.

From a sample size estimation perspective, the departure point for sample size estimation under the regression model is much the same as in the classical sample size models considered earlier in the chapter. We want to learn about program impact, as captured through our estimate $\hat{\beta}_1$ of β_1. Specifically, we wish to obtain a sample size estimate n large enough to be able to test the null and alternative hypotheses

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

with significance α and power $1 - \beta$. As we will see, doing this will, as before, require assumptions. And, as we will see, the more elaborate the model (and multiple regression is more elaborate than

[64] This is indeed a very common assumption. It is only really in the context of instrumental variables that heterogeneous treatment effects have been at the center of the impact evaluation estimation discussion.

simple comparison of means) the more assumptions that are required.

This is the big picture, and the reader will benefit from keeping sight of it. But the path to a sample size estimate n is, as we will see, a somewhat complicated one. The first step is to introduce some preliminaries.

Our departure point is the linear equation for the observed outcome variable Y in terms of P and X derived earlier through the partial observability potential outcomes framework:

$$Y_i = \beta_0 + \beta_1 \cdot P_i + \beta_2 \cdot X_i + \varepsilon_i$$

where i can index either individuals in a sample of size n or population of size N. A requirement for the linear regression model (and one imposed by the implicit "moment condition" behind ordinary least squares regression) is that $E(\varepsilon) = 0$. X can be thought of as containing one or morecontrol variables (say $X_{1i}, X_{2i}, ..., X_{ki}$ for $k \geq 2$). In other words, the variable X can represent either one or a vector of control variables.[65]

Ordinary least squares (OLS) regression of this model seeks to estimate the coefficients that minimize the sum of squares of the differences between the observed and predicted outcome values, i.e., residuals. That is, OLS regression finds the coefficient estimates $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ that minimize the sum squared residuals

$$\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} \hat{\varepsilon}_i^2$$

where

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot P_i + \hat{\beta}_2 \cdot X_i$$

for given data (Y_i, P_i, X_i) from a sample of n individuals indexed by i for $i = 1, ..., n$. Intuitively, OLS seeks estimates $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ that maximize the variation in Y_i, explained by the variation in the observed covariates (called regressors in this context) P_i and X_i.

But do those covariates explain any variation in Y_i? Unless the regressors are perfectly correlated, OLS will generate *some* estimate of $\hat{\beta}_1$ and $\hat{\beta}_2$. But does that mean that variation P_i and X_i predicts variation in Y_i in a statistically significant fashion?

To answer this question, one would perform a statistical test of the significance of the coefficient estimates $\hat{\beta}_1$ and $\hat{\beta}_2$. To perform the test, one would form a test statistic under the F-distribution. Before proceeding it might be worth briefly reviewing the F-distribution (with features like multiple degrees of freedom parameters, it can seem a bit unapproachable). In linear regression, the Student's t and normal distributions (leading to, respectively, t- and z-statistics) are typically employed for

[65] In the latter case, the linear equation for the outcome variable Y in terms of the program variable P and a set of control variables $\{X_1, ..., X_k\}$ can be defined as

$$Y_i = \beta_0 + \beta_1 \cdot P_i + \beta_{21} \cdot X_{1i} + \beta_{22} \cdot X_{2i} + ... + \beta_{2k} \cdot X_{ki} + \varepsilon_i$$

where we still impose $E(\varepsilon) = 0$. When we think of X_i as containing more than one control, $\beta_2 \cdot X_i$ is the vector inner product of a $1 \times k$ coefficient matrix β_2 and a $k \times 1$ vector of controls X_i (here we forego the niceties of indicating transposition, as in the oft adopted approach of assuming that both are $k \times 1$ and writing $\beta_2^T X_i$ where T indicates "transpose" and "·" is usually foregone in matrix multiplication settings).

tests of single parameter (such as β_1). For instance, form a test statistic under the null hypothesis

$$H_0 : \beta_1 = 0$$

and test assuming a t- or z-statistic (we would typically reject this null if the absolute value of the test statistic was large enough to make it improbable under the null hypothesis).

The F-distribution is a common choice for testing hypotheses regarding linear combinations of parameters in a linear model. As such, the F-distribution can be used to test single constraints such as

$$H_0 : \beta_1 = \beta_2$$

or multiple constraints such as[66]

$$H_0 : \beta_1 = \beta_2 = 0$$

The F-distribution will play a central role in this subsection since sample size estimation will depend on multiple covariates.

A variable v that follows the F-distribution with v_1 and v_2 degrees of freedom (denoted $v \sim F(v_1, v_2)$) has the probability density function

$$f\left(v; v_1, v_2\right) = \frac{\sqrt{\frac{(v_1 \cdot v)^{v}_{1} \cdot {v_2}^{v_2}}{(v_1 \cdot v + v_2)^{v_1 + v_2}}}}{v \cdot B\left(\frac{v_1}{2}, \frac{v_2}{2}\right)}$$

where $B(\cdot)$ is the beta function. The cumulative distribution is not much prettier. An F-distributed random variable can be characterized variously as the square of a Student's t-distributed random variable.

In terms of what the F-distribution actually looks like, we graph the F-density for several different combinations of degrees of freedom parameters in Figure 3.15. To begin with, note that in all cases the support of the F-density[67] is defined only for positive values for the F-distributed random variable. This is perhaps not surprising when one considers that an F-distributed random variable can be characterized as the ratio of two independent chi-squared random variables.[68] The degrees of freedom combinations are meant generally to be reflective of the sort that we might actually encounter in the kind of applications of F tests that will shortly become our focus. As one can plainly see, the F-distribution's shape depends heavily on the degrees of freedom. In general, and as this figure illustrates, the first degrees of freedom (v_1, also referred to as the numerator degrees

[66] This is a case of multiple constrains because it actually tests that two conditions ($\beta_1 = 0$ and $\beta_2 = 0$) hold.

[67] The density of a random variable is those values for the random variable that have positive probability.

[68] $v \sim F(v_1, v_2)$ can be expressed in terms of two independent chi-squared random variables $Q_1 \sim \chi^2(d_1)$ and $Q_2 \sim \chi^2(d_2)$ as

$$v = \frac{Q_1/v_1}{Q_2/v_2}$$

where $\chi^2(v_j)$ denotes a chi-squared distribution with v_j degrees of freedom (for $j = 1, 2$), which in turn can be expressed as a sum of the squares of d_j independent and normally distributed random variables. Therefore, the chi-squared density is defined for positive values.

of freedom) tends to be particularly important as a determinant of the shape of the distribution.

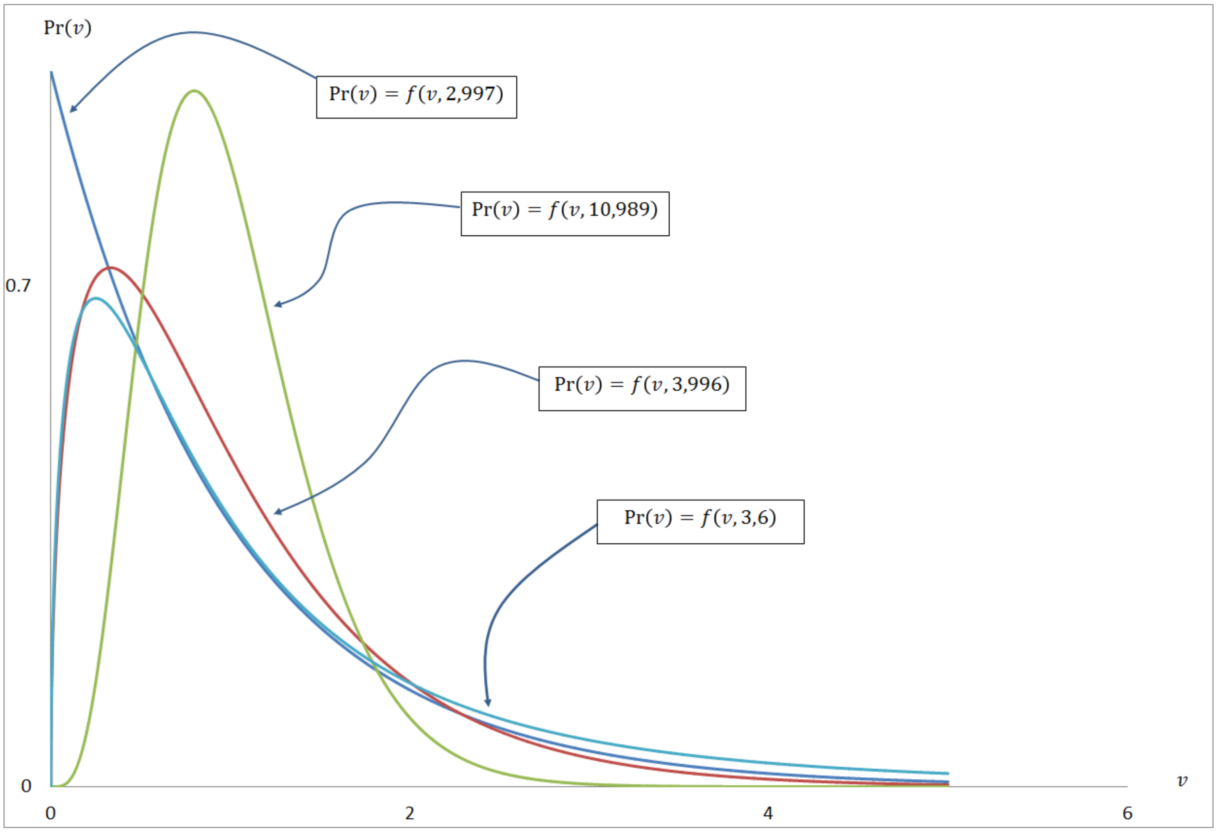


Figure 3.15. The F-distribution for various degrees of freedom

Within the regression context, the focus is usually on testing some restrictions on the basic general regression model

$$Y_i = \beta_0 + \beta_1 \cdot P_i + \beta_2 \cdot X_i + \varepsilon_i$$

This is the regression model adapted to our context, and X_i can be regarded as a vector of controls beyond program participation P_i. For the moment, we explicitly consider the case where we have a vector of k controls X_i, leading to

$$Y_i = \beta_0 + \beta_1 \cdot P_i + \beta_{21} \cdot X_{1i} + \beta_{22} \cdot X_{2i} + ... + \beta_{2k} \cdot X_{ki} + \varepsilon_i$$

Restrictions might be along the lines of (for example)

$$\beta_1 = \beta_{21}$$

$$\beta_{22} = 0$$

$$\beta_{24} + \beta_{25} + \beta_k = 30$$

One can think of two regression models, the unrestricted model that we have above

$$Y_i = \beta_0 + \beta_1 \cdot P_i + \beta_{21} \cdot X_{1i} + \beta_{22} \cdot X_{2i} + ... + \beta_{2k} \cdot X_{ki} + \varepsilon_i$$

and the model under all of these restrictions

$$Y_i = \beta_0 + \beta_1 \cdot P_i + \beta_1 \cdot X_{1i} + 0 \cdot X_{2i} + ... + (30 - \beta_{24} - \beta_{25}) \cdot X_{ki} + \varepsilon_i$$

The F-test would provide a joint test of all of these restrictions.

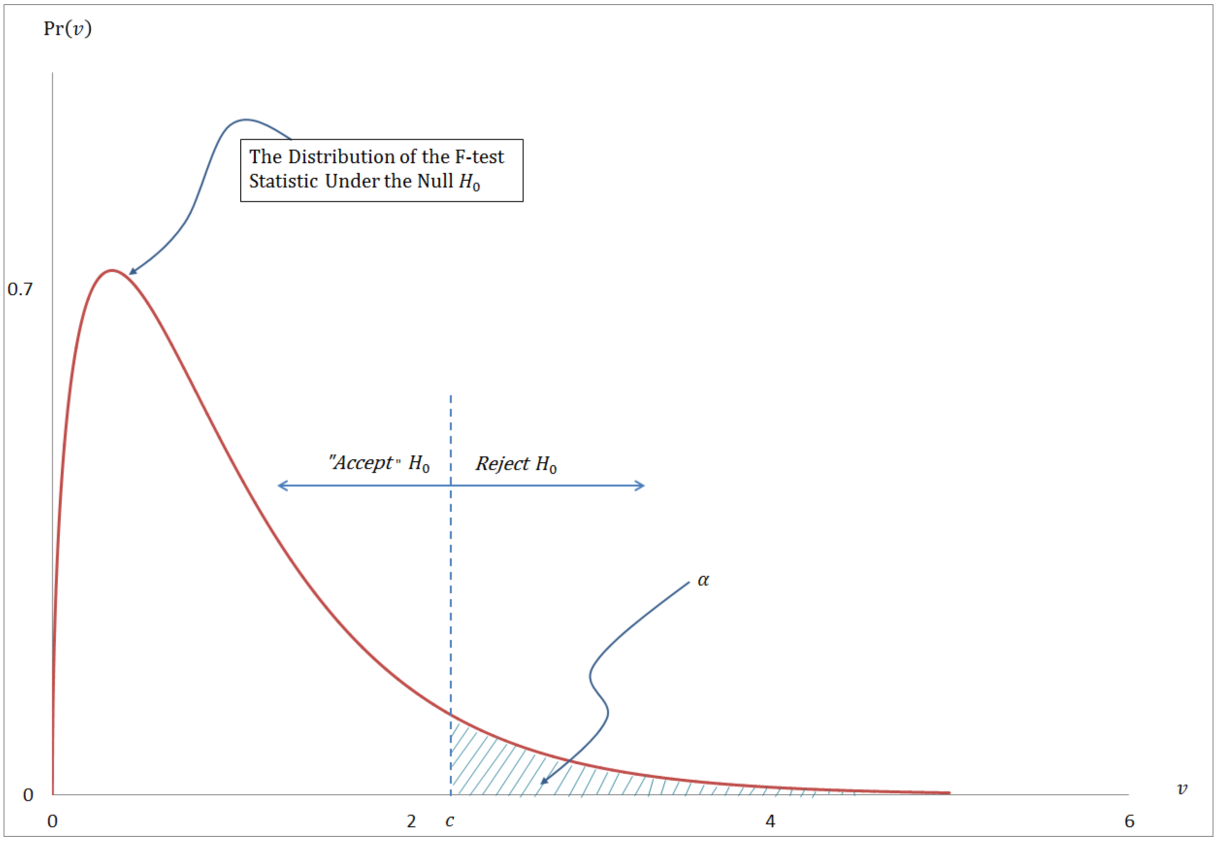


Figure 3.16. Acceptance and rejection with an F-test

Specifically, the null hypothesis would be

$$H_0 : \beta_1 = \beta_{21} \text{ and } \beta_{22} = 0 \text{ and } \beta_{24} + \beta_{25} + \beta_k = 30$$

against the alternative

$$H_a : \beta_1 \neq \beta_{21} \text{ or } \beta_{22} \neq 0 \text{ or } \beta_{24} + \beta_{25} + \beta_k \neq 30$$

We have seen the restricted and unrestricted regression models in the last paragraph. The sum of squared residuals for the restricted model would be

$$SS^{R}_{RESID} = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i^R\right)^2$$

while the same for the unrestricted model would be

$$SS^{U}_{RESID} = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i^U\right)^2$$

where $\hat{Y}_i^R$ is predicted Y for observation i using the estimated coefficients from the restricted model and $\hat{Y}_i^U$ is the same for the unrestricted model. Under the null hypothesis, the test statistic

$$f = \frac{\frac{SSR^{R}_{RESID} - SSR^{U}_{RESID}}{q}}{\frac{SSR^{U}_{RESID}}{(n-(k+1))}}$$

follows an F-distribution with q (where q is the number of restrictions being tested) and $(n-(k+1))$ (where n is the number of observations and k is the number of regressors in the unrestricted model) degrees of freedom.

With this test statistic f in hand, the logic of an F-test is pretty straightforward and should be familiar. Much as in the case of tests based on the Student's t and normal distributions, the test statistic f is formed under the null hypothesis. A critical value is determined such that the probability of a Type-I error is α. If the test statistic is larger than the critical value, the statistic has a value that is unlikely under the null and the null is hence rejected.

This is illustrated in Figure 3.16. Using the F-density with 3 and 996 degrees of freedom from Figure 3.15, Figure 3.16 shows the basic testing logic. The test statistic is constructed as if the null hypothesis H_0 is true. We choose a critical value (given the degrees of freedom) c from the F-distribution such that the probability of a Type-I error is α. If the test statistic exceeds this critical value c we reject H_0. If not we fail to reject the H_0.[69]

We now turn to the F-test with which we wish to begin our discussion of the preliminaries for sample size estimation. Specifically, we wish to test whether all of the regressors in the regression model determine Y. Specifically, we wish to test the null hypothesis

$$H_0 : \beta_1 = 0 \text{ and } \beta_2 = 0$$

or

$$H_0 : \beta_1 = \beta_2 = 0$$

[69] We select a value c arbitrarily; in Figure 3.16 it is slightly larger than 2. Further, per the discussion in the first section of this chapter, it is not really reasonable to say that we "accept" the null hypothesis H_0 since, in practice, we typically cannot differentiate between the null and slight deviations from it. We therefore do not say that we accept the null hypothesis H_0.

The unrestricted model is thus

$$Y_i = \beta_0 + \beta_1 \cdot P_i + \beta_2 \cdot X_i + \varepsilon_i$$

while the restricted model is

$$Y_i = \beta_0 + \varepsilon_i$$

In the event that the null hypothesis was true and the restricted model would then be appropriate, the estimand for the OLS estimate of β_0 would be $\overline{Y}$ (i.e., the population average of the observed outcome Y).

We have seen that one incarnation of the F statistic is

$$f = \frac{\frac{SS^R_{RESID} - SS^U_{RESID}}{q}}{\frac{SS^U_{RESID}}{(n-(k+1))}}$$

It turns out that there is a relatively easy way of stating this test statistic. First, we have[70]

$$\sum_{i=1}^{n} \left(Y_i - \hat{\overline{Y}}\right)^2 = \sum_{i=1}^{n} \left(\hat{Y}_i - \hat{\overline{Y}}\right)^2 + \sum_{i=1}^{n} \left(Y_i - \hat{Y}\right)^2$$

where $\hat{\overline{Y}}$ is an estimate of the population average of Y and

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot P_i + \hat{\beta}_2 \cdot X_i$$

is the OLS predicted value for Y for individual i ($\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ being the OLS estimates of the

[70] The proof of this is pretty straightforward

$$\sum_{i=1}^{n} \left(Y_i - \hat{\overline{Y}}\right)^2 = \sum_{i=1}^{n} \left(Y_i - \hat{\overline{Y}} + \hat{Y}_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} \left[\left(\hat{Y}_i - \hat{\overline{Y}}\right) + \left(Y_i - \hat{Y}_i\right)\right]^2 = \sum_{i=1}^{n} \left[\left(\hat{Y}_i - \hat{\overline{Y}}\right) + \left(\hat{\varepsilon}_i\right)\right]^2$$

$$= \sum_{i=1}^{n} \left[\left(\hat{Y}_i - \hat{\overline{Y}}\right)^2 + 2 \cdot \hat{\varepsilon}_i \cdot \left(\hat{Y}_i - \hat{\overline{Y}}\right) + \hat{\varepsilon}_i^2\right] = \sum_{i=1}^{n} \left[\left(\hat{Y}_i - \hat{\overline{Y}}\right)^2 + \sum_{i=1}^{n} \hat{\varepsilon}_i^2 + 2 \cdot \sum_{i=1}^{n} \hat{\varepsilon}_i \cdot \left(\hat{Y}_i - \hat{\overline{Y}}\right)\right]$$

$$= \sum_{i=1}^{n} \left[\left(\hat{Y}_i - \hat{\overline{Y}}\right)^2 + \sum_{i=1}^{n} \hat{\varepsilon}_i^2 + 2 \cdot \sum_{i=1}^{n} \hat{\varepsilon}_i \cdot \left(\hat{\beta}_0 + \hat{\beta}_1 \cdot P_i + \hat{\beta}_2 \cdot X_i - \hat{\overline{Y}}\right)\right]$$

$$= \sum_{i=1}^{n} \left(\hat{Y}_i - \hat{\overline{Y}}\right)^2 + \sum_{i=1}^{n} \hat{\varepsilon}_i^2 + 2 \cdot \left(\hat{\beta}_0 - \hat{\overline{Y}}\right) \cdot \sum_{i=1}^{n} \hat{\varepsilon}_i + 2 \cdot \hat{\beta}_1 \cdot \sum_{i=1}^{n} \hat{\varepsilon}_i \cdot P_i + 2 \cdot \hat{\beta}_2 \cdot \sum_{i=1}^{n} \hat{\varepsilon}_i \cdot X_i$$

The last three terms are all equal to zero per the moment conditions that OLS estimates satisfy, leaving us with:

$$\sum_{i=1}^{n} \left(Y_i - \hat{\overline{Y}}\right)^2 = \sum_{i=1}^{n} \left(\hat{Y}_i - \hat{\overline{Y}}\right)^2 + \sum_{i=1}^{n} \hat{\varepsilon}_i^2$$

thus establishing the proof.

regression model coefficients.) Notice that

$$\sum_{i=1}^{n}\left(Y_i-\hat{Y}\right)^2$$

is the predicted sum of squares that OLS seeks to minimize through its estimates of $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$. In other words

$$\sum_{i=1}^{n}\left(Y_i-\hat{Y}\right)^2=\sum_{i=1}^{n}\hat{\varepsilon}_i^2$$

We could thus write

$$\sum_{i=1}^{n}\left(Y_i-\hat{\bar{Y}}\right)^2=\sum_{i=1}^{n}\left(\hat{Y}_i-\hat{\bar{Y}}\right)^2+\sum_{i=1}^{n}\hat{\varepsilon}_i^2$$

What this says intuitively is that the sum of squared variation in Y around its mean can be decomposed into squared variation explained by the regression model

$$\sum_{i=1}^{n}\left(\hat{Y}_i-\hat{\bar{Y}}\right)^2$$

and squared variation not explained by the regression model

$$\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2=\sum_{i=1}^{n}\hat{\varepsilon}_i^2$$

Alternatively, some readers may have seen this expressed as "total sum of squares equals explained sum of squares plus residual sum of squares." This is also the basic decomposition of variance behind the analysis of variance (ANOVA).

The sum of squared residuals for the restricted model (i.e., the model under the null hypothesis $H_0:\beta_1=\beta_2=0$) is

$$\sum_{i=1}^{n}\left(Y_i-\hat{\bar{Y}}\right)^2$$

The sum of squared residuals under the unrestricted model is

$$\sum_{i=1}^{n}\left(Y_i-\hat{Y}\right)^2=\sum_{i=1}^{n}\hat{\varepsilon}_i^2$$

Per

$$\sum_{i=1}^{n}\left(Y_i-\hat{\bar{Y}}\right)^2=\sum_{i=1}^{n}\left(\hat{Y}_i-\hat{\bar{Y}}\right)^2+\sum_{i=1}^{n}\hat{\varepsilon}_i^2$$

the difference in the two is

$$\sum_{i=1}^{n}\left(Y_i-\hat{\bar{Y}}\right)^2-\sum_{i=1}^{n}\hat{\varepsilon}_i^2=\sum_{i=1}^{n}\left(\hat{Y}_i-\hat{\bar{Y}}\right)^2$$

Thus

$$f = \frac{\frac{SSR^R_{RESID} - SSR^U_{RESID}}{q}}{\frac{SSR^U_{RESID}}{(n-(k+1))}}$$

can be re-written

$$f = \frac{\frac{SSR^R_{RESID} - SSR^U_{RESID}}{q}}{\frac{SSR^U_{RESID}}{(n-(k+1))}} = \frac{\frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \hat{\bar{Y}}\right)^2}{q}}{\frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{(n-(k+1))}} = \frac{\frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \hat{\bar{Y}}\right)^2}{q}}{\frac{\sum_{i=1}^{n}\hat{\varepsilon}_i^2}{(n-(k+1))}}$$

The test statistic can thus be seen as involving a ratio of the explained sum of squares and the residual sum of squares for the unrestricted model. Eschewing explicit notation for restricted and unrestricted (since the unrestricted model is the basis of the terms that follow), it can be more compactly written as

$$f = \frac{\frac{SS_{REG}}{q}}{\frac{SS_{RESID}}{(n-(k+1))}}$$

where

$$SS_{REG} = \sum_{i=1}^{n}\left(\hat{Y}_i - \hat{\bar{Y}}\right)^2$$

and

$$SS_{RESID} = \sum_{i=1}^{n}\hat{\varepsilon}_i^2$$

where "REG" and "RESID" refer, respectively, to regression and residual.

This has been a lot of math to this point, and we thank the reader for their patience. For those already familiar with these results, this may have seemed a tediously redundant journey. However, it is important to have a solid sense of the foundations from which the sample size estimator is developed, and for that reason we have taken a rather exhaustive approach to this point. Later, the mathematical exposition will grow a bit breezier (in the sense of less explicit derivation of all results) in part because the derivations themselves might grow too involved, but the abbreviated discussion should be far easier to follow from this foundation.

As for the test statistic

$$f = \frac{\frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \hat{\bar{Y}}\right)^2}{q}}{\frac{\sum_{i=1}^{n}\hat{\varepsilon}_i^2}{(n-(k+1))}}$$

we can see a real intuition emerging in this formulation. This is the test statistic with q and $n-(k+1)$ degrees of freedom under the null hypothesis

$$H_0 : \beta_1 = \beta_2$$

We reject the null if this statistic is too large in value. That would happen if the explained sum of squares from the unrestricted model

$$\sum_{i=1}^{n}\left(\hat{Y}_i-\hat{\overline{Y}}\right)^2$$

was large compared with the residual sum of squares from that model

$$\sum_{i=1}^{n}\hat{\varepsilon}_i^2$$

In other words, the test statistic would be large, and we would reject the null hypothesis, if the unrestricted model explained a lot of variation in Y. This is most likely to be the case if P and X are significant determinants of variation in Y, which would mean that $\beta_1 \neq 0$, $\beta_2 \neq 0$ or both.[71]

This test is a fairly standard element of the OLS regression output for most commercial statistical packages. The STATA Output 3.3 (based on the STATA .do file fdensity.do) provides an example of standard OLS output for the STATA statistical package. This output is based on a simulated (a fancy way of saying "made up") example from fdensity.do. The test statistic for the null hypothesis that all regressor coefficients are zero is "`F(2, 997) = 15.05`." The degrees of freedom are 2 (which equals the number of regressors and hence restrictions being tested) and 997 (which is the number of observations, 1,000, minus the number of regressors (2) plus 1 (to account for the constant). The p-value for the test (i.e., the probability of the test statistic exceeding 15.05 if the null is true) is 0 (from "`Prob > F= 0.0000`"). This means that the null hypothesis that the regressor coefficients are all zero is rejected (and rather strongly).

We then go on to derive the test statistic (which is more precisely 15.054416) in two ways: from the retained calculations from the STATA OLS computations (from various elements of the vector `e(·)`) and through brute force calculation ourselves of the model and residual sum of squares.

STATA Output 3.3 (fdensity.do)

```
. reg y x1 x2

      Source |       SS           df       MS      Number of obs   =     1,000
-------------+----------------------------------   F(2, 997)       =     15.05
       Model |  778.576063         2  389.288032   Prob > F        =    0.0000
    Residual |  25781.1502       997  25.8587264   R-squared       =    0.0293
-------------+----------------------------------   Adj R-squared   =    0.0274
       Total |  26559.7263       999  26.5863126   Root MSE        =    5.0851

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |  -.7604283   .1590645    -4.78   0.000    -1.072568   -.4482888
          x2 |   .4919017   .1615362     3.05   0.002     .1749117    .8088918
       _cons |   1.347098    .161252     8.35   0.000     1.030666     1.66353
------------------------------------------------------------------------------

. * the (already reported) F-statistic
. di e(F)
15.054416
```

[71] Recall that the null hypothesis is that $\beta_1 = 0$ and $\beta_2 = 0$. Hence it would not fail if either $\beta_1 \neq 0$ or $\beta_2 \neq 0$.

```
. * where STATA stores the model sum of squares
. di e(mss)
778.57606

. * where STATA stores the residual sum of squares
. di e(rss)
25781.15

. * where STATA stores the model degrees of freedom
. di e(df_m)
2

. * where STATA stores the residual degrees of freedom
. di e(df_r)
997

. * Cranking out the F statistic the long way
. di (e(mss)/(e(df_m)))/(e(rss)/(e(df_r)))
15.054416

. di (e(mss)/2)/(e(rss)/(e(N)-2-1))
15.054416

. *Cranking our the F statistic the really long way
. predict yhat
(option xb assumed; fitted values)

. su y

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
           y |      1,000    1.400934    5.156192  -15.12218   18.22197

. loc a=r(mean)

. g SSREG=(yhat-`a´)^2

. g SSRRES=(y-yhat)^2

. total SSREG SSRRES

Total estimation                  Number of obs   =      1,000

--------------------------------------------------------------
             |      Total   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
       SSREG |   778.5761   36.34732      707.2502    849.9019
      SSRRES |   25781.15   1138.401      23547.22    28015.08
--------------------------------------------------------------

. loc b=_b[SSREG]

. loc c=_b[SSRRES]

. di `b´
778.57607

. di `c´
25781.15

. di (`b´/2)/(`c´/(e(N)-2-1))
15.054416
```

Having built this foundation of understanding of the F-test, we now introduce four measures of program impact that one could conceivably capture through the regression model: a multiple correlation coefficient, a partial correlation coefficient, a semi-partial correlation coefficient, and a regression coefficient. These measures relate to each other and can be recovered from one another when other necessary parameters are available. That is, one can formulate statistical hypothesis tests in terms of different measures based on available parameters in order to assess the same statistical hypothesis. We will overview the measures, their relationships, and hypothesis testing (specifically, how F-tests can be formulated).

3.3.1.1.1 Multiple Correlation Coefficient

We now discuss a measure of the overall explanatory performance of a regression model commonly referred to as the coefficient of determination but also referred to by some and in some contexts as the multiple correlation coefficient. It is often referred to simply as "R-squared," usually written R^2. We used the term multiple correlation coefficient because it is more consistent and compatible with the terminology employed for some of the other measures to follow. Indeed, this entire subsubsection is in some sense meant to ease into what follows and create a bridge to the earlier discussion of decomposition of variance and F-testing.

The coefficient of determination "R-squared" (R^2), a squared multiple correlation coefficient, is a measure of how much of the variation in Y is explained by the predicted Y, $\hat{Y}$, "fitted," or estimated OLS regression model

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot P_i + \hat{\beta}_2 \cdot X_i$$

It is

$$\hat{R}^2 = \frac{\sum_{i=1}^n \left(\hat{Y}_i - \hat{\bar{Y}}\right)^2}{\sum_{i=1}^n \left(Y_i - \hat{\bar{Y}}\right)^2} = 1 - \frac{\sum_{i=1}^n \left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^n \left(Y_i - \hat{\bar{Y}}\right)^2}$$

We place a hat "ˆ" over R^2 because it is, as a function of estimates, an estimate itself. The last term in the equality re-poses the explained proportion of variation as one minus the proportion not explained. Rather than use the conventional symbol, R^2, we will use the somewhat more elaborate

$$\hat{R}^2_{Y(P,X)} = \frac{\sum_{i=1}^n \left(\hat{Y}_i - \hat{\bar{Y}}\right)^2}{\sum_{i=1}^n \left(Y_i - \hat{\bar{Y}}\right)^2}$$

This is once again about alignment with the discussion to follow. The subscript $Y(P, X)$ is meant to indicate variation explained by the regressors P and X. We can also characterize this measure using the sum of squares"SS" notation introduced earlier:

$$\hat{R}^2_{Y(P,X)} = \frac{\sum_{i=1}^n \left(\hat{Y}_i - \hat{\bar{Y}}\right)^2}{\sum_{i=1}^n \left(Y_i - \hat{\bar{Y}}\right)^2} = \frac{SS_{REG[Y(P,X)]}}{SS_{TOTAL[Y]}}$$

where we now write "$SS_{REG[Y(P,X)]}$" to indicate that we are talking about the explained sum of squares

$$\sum_{i=1}^{n}\left(\hat{Y}_i - \hat{\bar{Y}}\right)^2$$

for variation in Y dependent on P and X and

$$SS_{TOTAL[Y]} = \sum_{i=1}^{n}\left(Y_i - \hat{\bar{Y}}\right)^2 = SS_{REG[Y(P,X)]} + SS_{RESID[Y(P,X)]}$$

Notice that we have not introduced the "$Y(P, X)$" qualifier to the subscript of $SS_{TOTAL[Y]}$. The reason for this is that the total variation is not dependent on our regression modelling approach. We have already established that

$$SS_{TOTAL[Y]} = \sum_{i=1}^{n}\left(Y_i - \hat{\bar{Y}}\right)^2 = SS_{REG[Y(P,X)]} + SS_{RESID[Y(P,X)]}$$

by a proof in an earlier footnote.

In STATA Output 3.4 we illustrate this measure of explained variation with the same STATA example from Output 3.3. First, we manually compute the total sum of squares. Then, we retrieve the explained sum of squares

$$\sum_{i=1}^{n}\left(\hat{Y}_i - \hat{\bar{Y}}\right)^2$$

from the STATA regression output parameter `e(mss)` (where the "m" in "mss" refers to *model*, as in "model sum of squares"). Dividing the explained/model sum of squares by the total sum of squares yields the multiple correlation coefficient. We then demonstrate that the same result can be achieved by subtracting from one the estimated residuals (stored post-estimation in `e(rss)`) divided by the total sum of squares.

STATA Output 3.4 (fdensity.do)

```
. * calculating total sum of squares (SS_TOTAL)
. su y

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
           y |      1,000    1.400934    5.156192  -15.12218   18.22197

. loc a=r(mean)

. g tot0=(y-`a´)^2

. total tot0

Total estimation                  Number of obs   =      1,000

--------------------------------------------------------------
             |      Total   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
        tot0 |   26559.73   1173.162      24257.58    28861.87
--------------------------------------------------------------

. loc totv=_b[tot0]
```

```
.
. * running the regression
.
. reg y x1 x2

      Source |       SS           df       MS      Number of obs   =     1,000
-------------+----------------------------------   F(2, 997)       =     15.05
       Model |  778.576063         2  389.288032   Prob > F        =    0.0000
    Residual |  25781.1502       997  25.8587264   R-squared       =    0.0293
-------------+----------------------------------   Adj R-squared   =    0.0274
       Total |  26559.7263       999  26.5863126   Root MSE        =    5.0851

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |  -.7604283   .1590645    -4.78   0.000    -1.072568   -.4482888
          x2 |   .4919017   .1615362     3.05   0.002     .1749117    .8088918
       _cons |   1.347098    .161252     8.35   0.000     1.030666     1.66353
------------------------------------------------------------------------------

. di e(r2)
.02931416

. di e(mss)
778.57606

. di e(rss)
25781.15

. * multiple correlation coefficient/coefficient of determination two ways
. di e(mss)/`totv´
.02931416

. di 1-e(rss)/`totv´
.02931416
```

The F-test of the null hypothesis

$$H_0 : \beta_1 = \beta_2 = 0$$

against

$$H_a : \beta_1 \neq 0 \text{ or } \beta_2 \neq 0 \text{ or both}$$

that we developed earlier can be translated into multiple correlation coefficient terms. First, in hypothesis terms, we can recast the null and alternative hypotheses as

$$H_0 : R^2_{Y(P,X)} = 0$$

$$H_a : R^2_{Y(P,X)} \neq 0$$

Since the variation captured or measured by $R^2_{Y(P,X)}$ is generated by P and X, these are equivalent ways of expressing the hypothesis (for instance, the only way that $R^2_{Y(P,X)}$ can equal zero if either

β_1 or β_2 or both do).[72]

The test statistic under this null hypothesis is

$$f = \frac{SS_{REG[Y(P,X)]}/2}{SS_{RESID[Y(P,X)]}/\left(n-3\right)} = \frac{\frac{\sum_{i=1}^{n}\left(\hat{Y}_i-\hat{\bar{Y}}\right)^2}{2}}{\frac{\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2}{n-3}}$$

$$= \frac{\frac{\frac{\sum_{i=1}^{n}\left(\hat{Y}_i-\hat{\bar{Y}}\right)^2}{\sum_{i=1}^{n}\left(Y_i-\hat{\bar{Y}}\right)^2}}{2}}{\frac{\frac{\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2}{\sum_{i=1}^{n}\left(Y_i-\hat{\bar{Y}}\right)^2}}{n-3}}$$

$$= \frac{R^2_{Y(P,X)}/2}{\left(1-R^2_{Y(P,X)}\right)/\left(n-3\right)}$$

Thus, the two conceptualizations of the test statistic are equivalent.

The null hypothesis is rejected at the significance level α when this test statistic is greater than c (in other words, $f > c$) where c represents the critical value from the F distribution

$$c = F_{2,n-3,1-\alpha}$$

In other words, the critical value is that at which the cumulative probability of an F-random variable with degrees of freedom 2 and $n-3$ is $1-\alpha$. α is thus the significance level of the test.

This test can easily be generalized to the case of more regressors (via either multiple program participation measures or multiple other controls X). Doing so requires (aside from estimating the regression model with the multiple regressors!) an adjustment to the degrees of freedom so that the numerator degrees of freedom equals the number of regressors and the denominator degrees of freedom equals $n-(k+1)$, where k is the number of regressors.

Notice that the test formulation in terms of a multiple correlation coefficient is appropriate when one seeks to test the overall contribution of all regressors in the model to explaining the variability in the outcome, Y. That is, it does not parse out the contribution of any specific regressors from other regressors in the model. In our example, we cannot test the contribution of the program P aside from the contribution of the control variable X. However, program evaluations are oftentimes interested

[72] Notice that the hypotheses are stated in terms of the population multiple correlation coefficient $R^2_{Y(P,X)}$ rather than the estimated multiple correlation coefficient $\hat{R}^2_{Y(P,X)}$. The population multiple correlation coefficient is

$$R^2_{Y(P,X)} = \frac{\sum_{i=1}^{N}\left(\tilde{Y}_i-\overline{Y}\right)^2}{\sum_{i=1}^{N}\left(Y_i-\overline{Y}\right)^2}$$

where $\overline{Y}$ is the population average for Y and

$$\tilde{Y}_i = \beta_0 + \beta_1 \cdot P_i + \beta_2 \cdot X_i$$

and N is the population size (which could be ∞).

in specific variables (such as the program variable P in our example) aside from control variables (for example, X). The other three measures that will be discussed in the following section (i.e., a partial correlation coefficient, a semi-partial correlation coefficient, and a regression coefficient) allow us to perform hypothesis tests focused on specific variables in the model.

3.3.1.1.2 Partial Correlation Coefficient

We now introduce an important twist on the multiple correlation coefficient. A squared partial correlation coefficient, denoted by $R^2_{Y(X_1).X_2}$, represents the amount of linear relationship between the outcome variable Y and the regressor(s) X_1 *after* accounting for linear relationship between Y and other regressor(s) (captured by X_2 in the expression $R^2_{Y(X_1).X_2}$).

Given the regression model setup to this point and the focus of this model our primary interest lies with $R^2_{Y(P).X}$. In other words, we are interested in the linear relationship between Y and P after accounting for the linear relationship between Y and other controls X. This can be expressed as

$$R^2_{Y(P).X} = \frac{SS_{REG[Y(P).X]}}{SS_{RESID[Y(X)]}}$$

where

$$SS_{REG[Y(P).X]} = SS_{REG[Y(P,X)]} - SS_{REG[Y(X)]}$$

In other words, the linear relationship between Y and P after controlling for the linear relationship between Y and X is the difference in the sum of squares from a regression of Y on P and X and from a regression of Y on X and the residuals from a regression of Y on X. The difference in the sum of squares from a regression of Y on P and X and from a regression of Y on X can be thought of as the additional explanatory capacity of program participation P. The residual sum of squares from the regression of Y on P can be thought of as the variation in Y not explained by X. $R^2_{Y(P).X}$ is thus the increment to explanatory power from program participation P (having already controlled for X) divided by the variation in Y not explained by X. It thus captures the fraction of variation in Y that is not explained by variation in X but is explained by variation in program participation P.

The statistical significance of $R^2_{Y(P).X}$ can be tested using an F-test. The null hypothesis in this case is that there is no linear association between the outcome variable Y and the regressor P after accounting for X. Most obviously, this can be stated

$$H_0 : R^2_{Y(P).X} = 0$$

$$H_a : R^2_{Y(P).X} \neq 0$$

A perhaps more revealing way of putting the null and alternative hypothesis would be

$$H_0 : \beta_1 = 0 | \beta_2$$

$$H_a : \beta_1 \neq 0|\beta_2$$

One interesting thing to note is that when $R^2_{Y(P,X)} = 0$, $R^2_{Y(P).X}$ is also 0. Because hypothesis testing based on a partial correlation coefficient is focused on the significance of a subset of regressors in the model, it is often called "nested hypothesis" testing.

The test statistic for this test is

$$f = \frac{\frac{SS_{REG[Y(P).X]}}{1}}{SS_{RESID[Y(P,X)]}/\left(n-3\right)} > c$$

The numerator degrees of freedom is 1 because, in this case, we are interested in only one regressor, program participation P, after controlling for X. Interestingly, the same hypothesis testing can be performed using the t-test because $t^2_{n-3} = F_{1,n-3}$, provided the alternative hypothesis is two-sided (in other words, $H_a : \beta_1 \neq 0|\beta_2$) and the numerator degrees of freedom is 1. In the more general case, where we are interested in multiple regressors (say, $\{P_1, ...P_q\}$, reflecting perhaps participation in q different components of a multi-component program or alternatively participation in q different programs) the numerator degrees of freedom would be q. That is, the concept of a partial correlation coefficient can be extended to test a linear relationship between Y and a set of q regressors for $q \geq 2$.

At significance level α, the null hypothesis of no linear association between the outcome variable Y and the program variable P after controlling for X is rejected if

$$f = \frac{SS_{REG[Y(P).X]}/1}{SS_{RESID[Y(P,X)]}/\left(n-3\right)} > c$$

where the critical value c is the value from the $F_{1,n-3,1-\alpha}$ distribution associated with a cumulative probability of $1-\alpha$. The significance of the test is thus α.

3.3.1.1.3 Semi-Partial Correlation Coefficient

A squared semi-partial correlation coefficient is another measure that represents the amount of linear relationship between the outcome variable Y and a set of q regressors of interest after controlling for other regressors in the model. A squared semi-partial correlation coefficient of the program variable P and the outcome variable Y controlling for X is denoted by $SR^2_{Y(P).X}$. It is closely related to a squared partial correlation coefficient, $R^2_{Y(P).X}$, and expressed as

$$SR^2_{Y(P).X} = R^2_{Y(P,X)} - R^2_{Y(X)} = \frac{SS_{REG[Y(P,X)]} - SS_{REG[Y(X)]}}{SS_{TOTAL[Y]}}$$

This is a very similar measure to $R^2_{Y(P).X}$. Notice that they have the same numerator, $SS_{REG[Y(P,X)]} - SS_{REG[Y(X)]}$, while they have a different denominator: $SS_{RESID[Y(X)]}$ for $R^2_{Y(P).X}$, and $SS_{TOTAL[Y]}$ for $SR^2_{Y(P).X}$. The difference in the denominator implies that

$$SR^2_{Y(P).X} < R^2_{Y(P).X}$$

under the same condition, because

$$SS_{TOTAL[Y]} > SS_{RESID[Y(X)]}$$

when $SS_{REG[Y(X)]} > 0$ (i.e., when Y is associated with X). Furthermore, $SR^2_{Y(P).X} = R^2_{Y(P).X}$ when $SS_{REG[Y(X)]} = 0$ (i.e., when Y and X are not associated with each other).

All of that being said, $SR^2_{Y(P).X}$ and $R^2_{Y(P).X}$ measure essentially the same thing: the amount of linear relationship between the outcome variable Y and regressor variables of interest after controlling for other regressor variables in the model. They are just formulated differently. To see this, a squared partial correlation coefficient can be re-written in terms of a squared semi-partial correlation coefficient as

$$\begin{aligned} R^2_{Y(P).X} &= SR^2_{Y(P).X} \cdot \frac{SS_{TOTAL[Y]}}{SS_{RESID[Y(X)]}} \\ &= SR^2_{Y(P).X} \cdot \frac{1}{\left(\frac{SS_{RESID[Y(X)]}}{SS_{TOTAL[Y]}}\right)} \\ &= SR^2_{Y(P).X} \cdot \frac{1}{\left(\frac{SS_{TOTAL[Y]}-SS_{REG[Y(X)]}}{SS_{TOTAL[Y]}}\right)} \\ &= SR^2_{Y(P).X} \cdot \frac{1}{1-R^2_{Y(X)}} \end{aligned}$$

and one can perform the same hypothesise testing in the case of the partial correlation coefficient.

3.3.1.1.4 Regression Coefficient

A regression coefficient is yet another (and perhaps the most direct and obvious!) measure of the linear relationship between the outcome variable Y and a regressor of interest after controlling for other regressors in the model. We will focus on two types of regression coefficients here: an unstandardized regression coefficient, and a standardized regression coefficient.

An *unstandardized* regression coefficient for P, denoted by β_1, represents the mean change in Y for one unit of change in P, accounting for X. This is the standard interpretation of the coefficient from straightforward OLS regression of one variable on another. A *standardized* regression coefficient is obtained from regressing standardized Y on standardized P and X. A standardized variable can be obtained by first subtracting its mean from the observed value and then dividing it by the standard deviation of the variable. To clarify, if w is a random variable, standardized w is

$$\frac{w-\overline{w}}{\sigma_w}$$

where $\overline{w}$ is the mean of w (ideally population but in practice we use the sample estimated mean $\hat{\overline{Y}}$) and σ_w is the standard deviation of w (again, ideally population but in practice we use the sample estimated standard deviation s_w). A standardized regression coefficient therefore represents the mean change in Y, measured in the number of standard deviations of Y, associated with a one standard deviation change in P while accounting for X. A standardized regression coefficient, which is denoted by $\beta_{s,1}$, can be expressed in terms of an unstandardized regression coefficient (β_1) as

$$|\beta_{s,1}| = \sqrt{\frac{SR^2_{Y(P).X}}{1 - R^2_{P(X)}}} = |\beta_1| \cdot \frac{\sigma_P}{\sigma_Y}$$

where σ_P and σ_Y denote the standard deviation of P and Y, respectively. There is clear correspondence between the two regression coefficients.[73]

The use of standardized regression coefficients is less common. Regression estimation typically proceeds with variables defined in their original units (i.e., with unstandardized variables). One of the major purposes of performing regression in the standardized unit is to make the regression coefficients "unit-free." Because standardized regression coefficients are expressed in terms of standard deviations of variables, they are not affected by the unit of the regressors or the outcome. This

[73] The proof of the equation is provided as follows. Let Y_s, P_s, and X_s denote standardized Y, P, and X, respectively. Let $\beta_{s,j}$ denote jth standardized regression coefficient for $j = 1, 2$ of the unstandardized regression coefficients, denoted as β_j, in $Y = \beta_0 + \beta_1 \cdot P + \beta_2 \cdot X + \varepsilon$. Recall that standardization of each variable can be obtained by first subtracting its mean from the observed value and then dividing it by the standard deviation of the variable.

$$Y = \beta_0 + \beta_1 \cdot P + \beta_2 \cdot X + \varepsilon$$

$$Y - \overline{Y} = [\beta_0 + \beta_1 \cdot P + \beta_2 \cdot X + \varepsilon] - \overline{Y} \text{ (subtracting } \overline{Y}\text{)}$$

$$\frac{Y - \overline{Y}}{\sigma_Y} = \frac{1}{\sigma_Y} \cdot \left[\beta_0 + \beta_1 \cdot P + \beta_2 \cdot X + \varepsilon - \overline{Y}\right] \text{ (deviding both sides by } \sigma_Y\text{)}$$

Notice that $\beta_0 = \overline{Y} - \beta_1 \cdot \overline{P} - \beta_2 \cdot \overline{X}$. Then by substituting for β_0 gives

$$\frac{Y - \overline{Y}}{\sigma_Y} = \frac{1}{\sigma_Y} \cdot \left[\left(\overline{Y} - \beta_1 \cdot \overline{P} - \beta_2 \cdot \overline{X}\right) + \beta_1 \cdot P + \beta_2 \cdot X + \varepsilon - \overline{Y}\right]. \text{ (substituting for } \beta_0 \text{ with } \left[\overline{Y} - \beta_1 \cdot \overline{P} - \beta_2 \cdot \overline{X}\right]\text{)}$$

$$\frac{Y - \overline{Y}}{S_Y} = \frac{1}{\sigma_Y} \cdot \beta_1 \cdot \left(P - \overline{P}\right) + \frac{1}{\sigma_Y} \cdot \beta_2 \cdot \left(X - \overline{X}\right) + \frac{\varepsilon}{\sigma_Y}. \text{ (rearranging the terms)}$$

$$\frac{Y - \overline{Y}}{\sigma_Y} = \beta_1 \cdot \frac{\sigma_P}{\sigma_Y} \cdot \frac{P - \overline{P}}{\sigma_P} + \beta_2 \cdot \frac{\sigma_X}{\sigma_Y} \cdot \frac{X - \overline{X}}{\sigma_X} + \frac{\varepsilon}{\sigma_Y}. \text{ (multiplying } \beta_1 \text{ and } \beta_2 \text{ by } \tfrac{\sigma_P}{\sigma_P} = 1 \text{ and } \tfrac{\sigma_X}{\sigma_X} = 1\text{, respectively)}$$

Given that $Y_s = \frac{Y - \overline{Y}}{\sigma_Y}$, $P_s = \frac{P - \overline{P}}{\sigma_P}$, and $X_s = \frac{X - \overline{X}}{\sigma_X}$, the equation can be rewritten as

$$Y_s = \beta_1 \cdot \frac{\sigma_P}{\sigma_Y} \cdot P_s + \beta_2 \cdot \frac{\sigma_X}{\sigma_Y} \cdot X_s + \frac{\varepsilon}{\sigma_Y},$$

and further

$$Y_s = \beta_{s,1} \cdot P_s + \beta_{s,2} \cdot X_s + \frac{\varepsilon}{\sigma_Y}.$$

where

$$\beta_{s,1} = \beta_1 \cdot \frac{\sigma_P}{\sigma_Y}$$
$$\beta_{s,2} = \beta_2 \cdot \frac{\sigma_X}{\sigma_Y}$$

thus establishing the proof. This implies that standardized regression coefficients can be recovered from their corresponding unstandardized regression coefficients by dividing them by the standard deviation of Y and multiplying by the standard deviation of their corresponding variable. In practice, the population means ($\overline{Y}$, $\overline{P}$, and $\overline{X}$) and the population standard deviations (σ_Y, σ_P, and σ_X) are unknown and we use the sample estimated means ($\hat{\overline{Y}}$, $\hat{\overline{P}}$, and $\hat{\overline{X}}$) and the sample estimated standard deviations (s_Y, s_P, and s_X) to estimate the standardized regression coefficients.

allows more direct comparison of the relative importance of the regressors in terms of explaining the variability in the outcome variable. That said, in program impact estimation we tend to think about and frame statistical hypothesis testing in terms of a mean change or mean difference in the outcome variable in its original unit instead of its standard deviation. Perhaps above all, it is generally not easy to communicate findings to the wider world in standardized form: funders, program officials, the media, etc. tend to think in terms of how much the act of program participation changes the level or probability of an event.

The hypothesis testing of the regression coefficient for P is again a nested hypothesis testing. That is, the null hypothesis (H_0) and alternative hypothesis (H_a) are formulated as

$$H_0 : \beta_1 = 0|\beta_2$$

$$H_a : \beta_1 \neq 0|\beta_2$$

This is statistically equivalent to formulating the hypotheses in terms of the squared partial correlation coefficient $R^2_{Y(P).X}$ or the squared semi-partial correlation coefficient $SR^2_{Y(P).X}$ as

$$H_0 : R^2_{Y(P).X} = 0$$

$$H_a : R^2_{Y(P).X} \neq 0$$

or

$$H_0 : SR^2_{Y(P).X} = 0$$

$$H_a : SR^2_{Y(P).X} \neq 0$$

This null hypothesis regarding the regression coefficient β_1 can be tested using the F-test. At significance level α, the null hypothesis of no linear association between the outcome variable Y and the program variable P after controlling for X is rejected if

$$f = \frac{SS_{REG[Y(P).X]}/1}{SS_{RESID[Y(P,X)]}/\left(n-3\right)} > c$$

where the critical value c is the value from the $F_{1,n-3,1-\alpha}$ distribution below which the cumulative probability is $1-\alpha$. Notice that the F-statistic is identical to those for the squared partial correlation coefficient ($R^2_{Y(P).X}$) or the squared semi-partial correlation coefficient ($SR^2_{Y(P).X}$) when the number of variables simultaneously tested (i.e., P in this case) is one.

Readers may be more used to seeing test results reported in terms of a t-, rather than F-, statistic in the standard regression output from commercial statistical software packages such as STATA. However, this is equivalent to the aforementioned hypothesis testing using the F-statistic. Though we have alluded to this before, we now briefly review the relationship between t and F distributions to understand the equivalence of hypothesis testing.

Let Q_1 and Q_2 be random variables with chi-squared distributions with degrees of freedom of

ν_1 and ν_2, respectively, i.e., $Q_1 \sim \chi^2(\nu_1)$ and $Q_2 \sim \chi^2(\nu_2)$. Further, assume that Q_1 and Q_2 are independently distributed. Let U be a random variable of a form $U = \frac{Q_1/\nu_1}{Q_2/\nu_2}$. Then $U \sim F(\nu_1, \nu_2)$, that is, U follows an F-distribution with the numerator degrees of freedom of ν_1 and the denominator degrees of freedom of ν_2. When $\nu_1 = 1$ (i.e., only one variable is tested simultaneously) U reduces to $U = \frac{Q_1}{Q_2/\nu_2}$ and $U \sim F(1, \nu_2)$. Now let W be a random variable defined as $W = \frac{Z}{\sqrt{Q_2/\nu_2}}$ where $Z \sim N(0,1)$ (i.e., has a standard normal distribution). Then W follows the t-distribution with ν_2 degrees of freedom, i.e., $W \sim t(\nu_2)$. Then $W^2 = \frac{Z^2}{Q_2/\nu_2}$. Because $Z^2 = \chi_1^2$ (i.e., has a χ^2 distribution with 1 degree of freedom), W^2 is equivalent to U when $\nu_1 = 1$, that is, $t^2(\nu_2) = F(1, \nu_2)$. Therefore, when performing hypothesis testing of one regression coefficient, which give the numerator degrees of freedom (ν_1) of 1, one can apply either an F-test or a t-test to obtain the same statistical results. Note that this exchangeability applies only to cases in which $\nu_1 = 1$. When testing multiple variables simultaneously, an F-test needs to be applied.

3.3.1.1.5 The Relationships between the Three Measures

Before moving on to discussing power and sample size calculations, we review the relationships among the three measures of linear relationship between the outcome variable Y and P after accounting for X: a squared partial correlation coefficient ($R^2_{Y(P).X}$), a squared semi-partial correlation coefficient ($SR^2_{Y(P).X}$), and a regression coefficient (β_1). A clear understanding of the three measures is helpful for stating hypothesis testing in different measures and proceeding with power and sample size calculations.

The unstandardized regression coefficient for P, denoted by β_1, can be expressed in terms of a squared semi-partial correlation coefficient as

$$|\beta_1| = \sqrt{\frac{SR^2_{Y(P).X}}{1 - R^2_{P(X)}}} \cdot \frac{\sigma_Y}{\sigma_P}$$

where σ_Y and σ_P are the standard deviation of Y and P, respectively. The term $1 - R^2_{P(X)}$ presents the amount of variation in P unexplained by X as

$$1 - R^2_{P(X)} = 1 - \frac{SS_{REG[P(X)]}}{SS_{TOTAL[P]}} = \frac{SS_{RESID[P(X)]}}{SS_{TOTAL[P]}}$$

which implies that the term $1 - R^2_{P(X)}$ measures the amount of "tolerance" of P for X. The higher the tolerance for P is (or in other words, the smaller the correlation between P and X is), the larger the amount of variability in Y that can be uniquely accounted for by P.

The squared partial correlation coefficient can be rewritten in terms of a squared semi-partial correlation coefficient as

$$R^2_{Y(P).X} = SR^2_{Y(P).X} \cdot \frac{SS_{TOTAL[Y]}}{SS_{RESID[Y(X)]}} = SR^2_{Y(P).X} \cdot \frac{1}{1 - R^2_{Y(X)}}$$

Thus, the three measures — a squared partial correlation coefficient ($R^2_{Y(P).X}$), a squared semi-partial correlation coefficient ($SR^2_{Y(P).X}$), and a regression coefficient (β_1) — can be recovered from one another once the other necessary parameters (such as $1 - R^2_{P(X)}$, σ_P, and σ_Y) are obtained for the transformation. Moreover, the fact that the three measures represents the same concept (i.e., the amount of linear relationship between the outcome variable Y and a regressor of interest after accounting for other regressors in the model) is reflected in the statistical hypotheses formulated in different measures while testing the same hypothesis.

3.3.1.1.6 Sample Size Estimation for OLS

The first section of this chapter was concerned primarily with the classic testing approach for comparing estimates of means across samples or an estimate of mean from a single sample against a theoretical value. A major lesson of the discussion (one that prompted the name of the chapter) is that such basic sample size estimation required a number of informed guesses in the form of assumptions. Some of these included: basic distributional assumptions against which to assess the value of the test statistic, assumptions about the true value of the population average under the null hypothesis, a variance of the outcome variable for each group,[74] etc. Moreover there were assumptions that had to be made about the basic parameters for the test. What minimum degree of difference from the null hypothesis did we wish to detect and with what power? What degree of significance did we want for the test?

Much of this carries over to sample size estimation for OLS, though the manner with which these factors enter into consideration is in some respects more complicated. Some of the key parameters for OLS sample size estimation depend on the sample size n (which, after all, we seek to determine). Because of this, sample size estimations for OLS cannot be solved through a closed-form formula. In other words, the more straightforward sample size estimation formulas from the first section of this chapter involved explicitly solving for n. This are no such OLS sample size estimation formulas of which we are aware. The OLS formulas thus solve for n by iteration to find a sample size n large enough to satisfy conditions of parameters that depend on a sample size. In the following section, we will examine the sample size estimation methods for OLS.

The foundation of power and sample size estimations for OLS was developed by Cohen (1977, 1988) and further extended by Gatsonis and Sampson (1989). An excellent review of the topic is provided by Maxwell (2000). Cohen (1977, 1988) provides guidance for power and sample size estimations for multiple linear regression treating regressors as fixed. In other words, Cohen does not treat the regressors as a source of random variation in Y but instead as essentially fixed quantities. Instead, the outcome variable Y is independently and normally distributed conditional on the predictors with a constant variance supplied exclusively via the error term. These assumptions enable sample size and power estimations relying on the F- (or t-) distribution, with hypothesis tests that are robust to moderate violations of these assumptions.

[74] In the case of tests regarding the population average of a continuous variable.

A general form of sample size can be written as a function of key parameters as

$$n = \frac{\lambda}{f_{idx}^2} + k - q$$

where k is the total number of regressors in the model and q is the number of regressors tested simultaneously. For our test of P after controlling for X, $k = 2$ and $q = 1$. f_{idx}^2 is defined as an "effect size index," introduced by Cohen (1977, 1988), and can be expressed as a function of correlation coefficients or regression coefficients.[75] Because correlation coefficients and regression coefficients capture the amount of linear relationship between Y and P relative to the residuals after accounting for X, f_{idx}^2 can be seen as a function of the degree to detect, d, and the variability in Y. We will discuss how to specify the value of f_{idx}^2 for hypotheses framed in different measures, namely $R^2_{Y(P,X)}$, $SR^2_{Y(P).X}$, and β_1.

First we demonstrate how to specify f_{idx}^2 for a hypothesis framed in the squared multiple correlation coefficient $R^2_{Y(P,X)}$. The effect size index f_{idx}^2 can be expressed in terms of $R^2_{Y(P,X)}$ as

$$f_{idx}^2 = \frac{R^2_{Y(P,X)}}{1 - R^2_{Y(P,X)}}$$

Recall that the test of the null hypothesis

$$H_0 : \beta_1 = \beta_2 = 0$$

against

$$H_a : \beta_1 \neq 0 \text{ or } \beta_2 \neq 0 \text{ or both}$$

is statistically equivalent to

$$H_0 : R^2_{Y(P,X)} = 0$$

against

$$H_a : R^2_{Y(P,X)} \neq 0$$

As we discussed in the first section of this chapter, we will need to frame the alternative hypothesis in terms of a more specific threshold for violation $R^2_{Y(P,X)} \neq 0$ that we wish to detect. This is accomplished by specifying the degree d_R to detect as

$$H_a : R^2_{Y(P,X)} \geq d_R$$

Then the value of f_{idx}^2 that we need for performing the hypothesis testing for a specified value of d_R is derived as

$$f_{idx}^2 = \frac{d_R}{1 - d_R}$$

[75] The effect size index as introduced by Cohen (1977, 1988) is denoted by f in the publications. We use the notation f_{idx} instead to avoid possible confusion with an F-statistic.

which is expressed in terms of d_R.

Next we demonstrate how to specify f^2_{idx} for a hypothesis framed in the squared semi-partial correlation coefficient $SR^2_{Y(P).X}$, the squared partial correlation coefficient $R^2_{Y(P).X}$, or β_1. The effect size index f^2_{idx} can be expressed in terms of $SR^2_{Y(P).X}$ or $R^2_{Y(P).X}$ as

$$f^2_{idx} = \frac{R^2_{Y(P,X)} - R^2_{Y(X)}}{1 - R^2_{Y(P,X)}} = \frac{SR^2_{Y(P).X}}{1 - R^2_{Y(P,X)}}$$

This definition is applicable to hypothesis testing focused on the regressor P after controlling for the regressor X as

$$H_0 : \beta_1 = 0|\beta_2$$

against

$$H_a : \beta_1 \neq 0|\beta_2$$

which is statistically equivalent to

$$H_0 : SR^2_{Y(P).X} = 0$$

against

$$H_a : SR^2_{Y(P).X} \neq 0$$

or

$$H_0 : R^2_{Y(P).X} = 0$$

against

$$H_a : R^2_{Y(P).X} \neq 0$$

Again, we will need to frame the alternative hypothesis in terms of a more specific threshold for violation $\beta_1 \neq 0|\beta_2$ and $SR^2_{Y(P).X} \neq 0$ that we wish to detect. This is accomplished by specifying the degree to detect for β_1, $SR^2_{Y(P).X}$, and $R^2_{Y(P).X}$, denoted by d_β, d_{SR}, and d_{PR}, respectively, as

$$H_a : \beta_1 \leq -d_{\beta_1}|\beta_2$$

$$H_a : \beta_1 \geq d_{\beta_1}|\beta_2$$

and

$$H_a : SR^2_{Y(P).X} \geq d_{SR}$$

and

$$H_a : R^2_{Y(P).X} \geq d_{PR}$$

Then f^2_{idx} for performing the hypothesis testing can be derived as

$$f^2_{idx} = \frac{d_{SR}}{1 - R^2_{Y(P,X)}}$$

which requires a prior for $R^2_{Y(P,X)}$.

We have reviewed that the three measures, namely, the squared partial correlation coefficient $R^2_{Y(P).X}$, squared semi-partial correlation coefficient $SR^2_{Y(P).X}$, and the regression coefficient β_1, can be recovered from one another once the other necessary parameters (for example, $1 - R^2_{P(X)}$, σ_P, and σ_Y) are obtained for the transformation. This suggests that f^2_{idx} can be specified in terms of $R^2_{Y(P).X}$ or β_1 as well.

Given that

$$R^2_{Y(P).X} = SR^2_{Y(P).X} \cdot \frac{1}{1 - R^2_{Y(X)}}$$

f^2_{idx} can be expressed in terms of a squared partial correlation coefficient as

$$f^2_{idx} = \frac{SR^2_{Y(P).X}}{1 - R^2_{Y(P,X)}} = \frac{R^2_{Y(P).X}}{1 - R^2_{Y(P).X}}$$

Then f^2_{idx} for performing the hypothesis testing for a specified value of d_{PR} is derived as

$$f^2_{idx} = \frac{d_{PR}}{1 - d_{PR}}$$

Likewise, given that

$$|\beta_1| = \sqrt{\frac{SR^2_{Y(P).X}}{1 - R^2_{P(X)}}} \cdot \frac{\sigma_Y}{\sigma_P}$$

f^2_{idx} can be expressed in terms of the regression coefficient for P (i.e., β_1) as[76]

$$f^2_{idx} = \frac{\beta_1^2 \cdot \left(1 - R^2_{P(X)}\right) \cdot \frac{\sigma_P}{\sigma_Y}}{1 - R^2_{Y(P,X)}}$$

Then f^2_{idx} for performing the hypothesis testing for a specified value of d_β is derived as

$$f^2_{idx} = \frac{d_\beta^2 \cdot \left(1 - R^2_{P(X)}\right) \cdot \frac{\sigma_P}{\sigma_Y}}{1 - R^2_{Y(P,X)}}$$

which requires priors for $R^2_{P(X)}$, σ_P, and σ_Y.

Next, let us examine λ. λ represents the noncentrality parameter for a noncentral F-distribution,

[76] Additionally, f^2_{idx} can be expressed in terms of the standardized regression coefficient for P (i.e., $\beta_{s,1}$) as

$$f^2_{idx} = \frac{\beta^2_{s,1} \cdot \left(1 - R^2_{P(X)}\right)}{1 - R^2_{Y(P,X)}}$$

Then for a specified value of degree to detect d_{β_s} for $\beta_{s,1}$, f^2_{idx} is derived as

$$f^2_{idx} = \frac{d^2_{\beta_s} \cdot \left(1 - R^2_{P(X)}\right)}{1 - R^2_{Y(P,X)}}$$

which requires priors for $R^2_{P(X)}$ and $R^2_{Y(P,X)}$.

which is relevant under the alternative hypothesis and hence speaks to power. Recall that hypotheses tests of correlation coefficients and regression coefficients can be performed based on an F-statistic (or a t-statistic when the numerator degrees of freedom is 1). When the null hypothesis is true, the test statistics for testing the correlation coefficients or regression coefficients follows an F-distribution. This is the reason the critical value for each test is obtained from an F-distribution with corresponding degrees of freedom. When the null hypothesis is false, the test statistics follow a noncentral F-distribution, whose shape is determined by the noncentrality parameter λ in addition to numerator and denominator degrees of freedom (ν_1, and ν_2). Specifically, the statistical power can be computed as

$$Power = 1 - \beta = \Pr\left(F_{\nu_1,\nu_2,\lambda} \geq F_{\nu_1,\nu_2,1-\alpha}\right)$$

where $F_{\nu_1,\nu_2,\lambda}$ is derived from the noncentral F-distribution with the numerator degrees of freedom ν_1 (i.e., the number of regressors that are tested simultaneously, q), the denominator degrees of freedom ν_2 (i.e., the sample size minus 1 and the number of all regressors, $n-(k+1)$), and the noncentrality parameter λ. For a fixed set of α, ν_1, and ν_2, a larger λ provides a larger statistical power (and vice versa). That is, λ can be seen as a measure of the difference in β_1 between false null hypothesis and true null hypothesis.

In other words, for a given level of statistical power and significance, λ is a function of degrees of freedom $\nu_1 = q$ and $\nu_2 = n-(k+1)$. Because $\nu_2 = n-(k+1)$ is a function of sample size, which we seek to determine, the sample size is not obtainable from a closed-form formula (i.e., one for which we can analytically solve for sample size n, as in the sample size formulas of the first section). Solving for the sample size thus requires iteration.

The iterative process runs roughly as follows:

1. Adopt an initial value for the sample size n.

2. We then obtain
$$\lambda = f^2_{idx} \cdot (\nu_1 + \nu_2 + 1) = f^2_{idx} \cdot (n - k + q)$$

3. Using the obtained λ, we calculate statistical power as $\Pr\left(F_{\nu_1,\nu_2,\lambda} \geq F_{\nu_1,\nu_2,1-\alpha}\right)$.

4. If the statistical power is sufficient we stop. If not we increase n (most plausibly by 1)[77] and return to step to 2. and repeat from there.

As iterative solution processes go, this is fairly straightforward. Cohen (1977, 1988) provides a set of tables of λ for combinations of numerator and denominator degrees of freedom, significance level, and power to facilitate the iterative solution process without a computing capacity.

The guidance proposed by Cohen (1977, 1988) for sample size and power estimations assumes that the regressors in the model are fixed. The fixed regressor model was reviewed by Gatsonis and Sampson (1989) from the theoretical point of view of conditional and unconditional models, that is, those involving fixed and random regressors, respectively. Unconditional models assume

[77] Though some algorithms may be more sophisticated. For instance, the increase in n might be larger if the power falls far short of the target. The increments to n might then fall as power got closer to the target.

that predictors in the model are random, and an outcome variable Y and a set of predictors in the model have a joint multivariate normal distribution. Their sample size and power estimations need to take into account the additional variability due to the random predictors, which generally requires a larger sample size compared to the same model with fixed regressors. Gatsonis and Sampson (1989) argued that an unconditional model should be employed in studies when the values of regressors cannot be controlled by the study design, which is typically the case in non-experimental, observational studies. Additionally, they demonstrated that Cohen's method, which builds on fixed regressor models, is a valid approximation to models based on random regressors and suggested simple modifications to Cohen's estimated statistical power to account for additional variability in the observed outcome variable arising from random regressors.[78] The topic is beyond the scope of the manual and an interested reader is referred to Gatsonis and Sampson (1989).

A number of modifications or extensions have been proposed to build on the foundational work by Cohen (1977, 1988) and Gatsonis and Sampson (1989). Specifically, many have attempted to address the methodological difficulty of carrying out the power or sample size estimation based on a multiple (partial) correlation coefficient (Dupont & Plummer, 1998; Kelly & Maxwell, 2003). Maxwell (2000) provides an excellent review of different approaches to power and sample size estimations for multiple linear regression. However, the method proposed by Cohen (1977, 1988) has been most commonly employed.

3.3.1.1.7 Sample Size Estimation for OLS in STATA

In this section we demonstrate how to perform sample size estimation in STATA for the multiple linear regression model of

$$Y = \beta_0 + \beta_1 \cdot P + \beta_2 \cdot X + \varepsilon$$

Our aim is to perform hypothesis testing on β_1 controlling for β_2 as

$$H_0 : \beta_1 = 0|\beta_2$$

$$H_a : \beta_1 \neq 0|\beta_2$$

which are statistically equivalent to

$$H_0 : R^2_{Y(P).X} = 0$$

$$H_a : R^2_{Y(P).X} \neq 0,$$

$$H_0 : SR^2_{Y(P).X} = 0$$

[78] While conditional and unconditional models have different model assumptions about their predictors, the usual test statistics that they rely on have the same null distribution while the alternative distribution differs, which leads to the specification of the statistical power as

$$Power = 1 - \beta = \Pr\left(\frac{\chi^2_{k+2\Theta}/q}{\chi^2_{n-k-1}/(n-k-1)} \geq F_{1-\alpha,q,n-k-1}\right)$$

where Θ is a random variable with a negative binomial distribution, k is the total number of regressors in the model, and q is the number of variables to test (i.e., the number of program variables).

$$H_a : SR^2_{Y(P).X} \neq 0,$$

and

$$H_0 : \beta_{s,1} = 0|\beta_{s,2}$$

$$H_a : \beta_{s,1} \neq 0|\beta_{s,2}$$

As we discussed in the previous section, we will need to frame the alternative hypothesis in terms of a more specific threshold for the violation that we wish to detect, i.e., d_β, as

$$H_a : \beta_1 \geq \ d_\beta|\beta_2 \text{ or } \beta_1 \leq \ -d_\beta|\beta_2$$

and equivalently, for hypotheses framed in $R^2_{Y(P).X}$, $SR^2_{Y(P).X}$, and $\beta_{s,1}$

$$H_a : R^2_{Y(P).X} \geq d_{PR}$$

$$H_a : SR^2_{Y(P).X} \geq d_{SR}$$

and

$$H_a : \beta_{s,1} \geq \ d_{\beta_s}|\beta_{s,2} \text{ or } \beta_{s,1} \leq \ -d_{\beta_s}|\beta_{s,2}$$

where d_{PR}, d_{SR}, and d_{β_s} denote the degree to detect for the hypothesis testing of $R^2_{Y(P).X}$, $SR^2_{Y(P).X}$ and $\beta_{s,1}$, respectively, as defined in the previous section.

We base our discussion on these hypotheses in this section, which is structured as follows: We first introduce the STATA command `powerreg`, which can perform power and sample size estimations for OLS, and review necessary parameter specifications to use the command. We then demonstrate the procedure of sample size estimation through numerical examples. We discuss the effect of different parameter specifications on estimated sample size, and conclude the section by discussing how to obtain priors for required parameters.

3.3.1.1.8 STATA Command powerreg

Without a computing capacity, the iterative solution process as proposed by Cohen (1977, 1988) required manual computation by consulting the set of tables of λ for combinations of numerator and denominator degrees of freedom, significance level, and power. However, we benefit from the recent development in computing capacity and usability of statistical software, including STATA. In STATA, the sample size estimation for fixed models as proposed by Cohen (1977, 1988) can be performed using a user-written command `powerreg`.[79] The basic syntax for the command

[79] The program was written by Philip B. Ender and Xiao Chen at Statistical Computing and Consulting, UCLA Academic Technology Services, and can be downloaded and installed by using the findit command, `findit powerreg`. The Institute for Digital Research and Education, UCLA, provides on its website examples of how to use the `powerreg` command. See http://www.ats.ucla.edu/stat/stata/dae/powerreg.htm/. The `powerreg` program allows for both sample size and statistical power estimations. The program recognizes which output is required by reading which parameters are specified by a user and which parameter specification is left unspecified. When estimating power, a sample size should be specified in `n(value)` while power as in `power(value)` is left out of the command line. When estimating a sample size, a sample size `n(value)` should be left out of the command line.

`powerreg` is:

```
powerreg , r2f(value) r2r(value) [ nvar(value) ntest(value) n(value)
alpha(value) power(value) ]
```

where `r2f` represents the squared multiple correlation coefficient under the unrestricted model, $R^2_{Y(P,X)}$; and `r2r` represents the squared multiple correlation coefficient under the restricted model, $R^2_{Y(X)}$. Notice that the difference between `r2f` and `r2r` ($R^2_{Y(P,X)}$ and $R^2_{Y(X)}$) is $SR^2_{Y(P).X}$, the squared semi-partial correlation coefficient of Y on P controlling for X. Then the program allows for specification of other parameters: the total number of regressors in the model (k) in `nvar`; the number of regressors tested simultaneously (q) in `ntest`; the sample size (n) in `n`; the statistical significance level (α [two-sided]) in `alpha`; and the statistical power level ($1-\beta$) in `power`. We specify a prior for each parameter in the bracket. For instance, a significance level of 0.05 is specified as `alpha(0.05)` in the command line.

One caveat of using `powerreg` is that `powerreg` requires users to frame hypotheses in terms of two squared multiple correlation coefficients, $R^2_{Y(P,X)}$ and $R^2_{Y(X)}$. Therefore when the hypotheses are framed in different parameters, that is, a regression coefficient (β_1 or $\beta_{s,1}$), a squared partial correlation coefficient ($R^2_{Y(P).X}$), or a squared semi-partial correlation coefficient($SR^2_{Y(P).X}$) to perform the same statistical testing, `powerreg` requires users to specify them in terms of two squared multiple correlation coefficients, $R^2_{Y(P,X)}$ and $R^2_{Y(X)}$. The first step in sample size estimation therefore is to reframe hypotheses framed in a parameter other than $R^2_{Y(P,X)}$ and $R^2_{Y(X)}$ (β_1, $\beta_{s,1}$, $R^2_{Y(P).X}$, or $SR^2_{Y(P).X}$) in terms of $R^2_{Y(P,X)}$ and $R^2_{Y(X)}$. While the transformation from either $R^2_{Y(P).X}$ or $SR^2_{Y(P).X}$ to $R^2_{Y(P,X)}$ and $R^2_{Y(X)}$ is rather simple and straightforward, the transformation from β_1 requires specification of additional parameters, namely $R^2_{P(X)}$, σ_Y, and σ_P.

We first demonstrate how to transform $SR^2_{Y(P).X}$ into $R^2_{Y(P,X)}$ and $R^2_{Y(X)}$. Recall that

$$SR^2_{Y(P).X} = R^2_{Y(P,X)} - R^2_{Y(X)}$$

Then $R^2_{Y(P,X)}$ and $R^2_{Y(X)}$ are derived as

$$R^2_{Y(P,X)} = SR^2_{Y(P).X} + R^2_{Y(X)} = d_{SR} + R^2_{Y(X)}$$

and

$$R^2_{Y(X)} = R^2_{Y(P,X)} - SR^2_{Y(P).X} = R^2_{Y(P,X)} - d_{SR}$$

suggesting that once a prior for either $R^2_{Y(P,X)}$ or $R^2_{Y(X)}$ is specified, the other can be derived for a specified d_{SR}.

We can transform $R^2_{Y(P).X}$ into $R^2_{Y(P,X)}$ and $R^2_{Y(X)}$ in a similar fashion. Notice that $R^2_{Y(P).X}$ can be expressed as

$$R^2_{Y(P).X} = SR^2_{Y(P).X} \cdot \frac{1}{1 - R^2_{Y(X)}}$$

$$= SR^2_{Y(P).X} \cdot \frac{1}{1 - R^2_{Y(P,X)} + SR^2_{Y(P).X}}$$

Then $R^2_{Y(P,X)}$ and $R^2_{Y(X)}$ can be derived as

$$R^2_{Y(P,X)} = 1 - \frac{SR^2_{Y(P).X}}{R^2_{Y(P).X}} + SR^2_{Y(P).X} = 1 - \frac{SR^2_{Y(P).X}}{d_{PR}} + SR^2_{Y(P).X}$$

and

$$R^2_{Y(X)} = 1 - \frac{SR^2_{Y(P).X}}{R^2_{Y(P).X}} = 1 - \frac{SR^2_{Y(P).X}}{d_{PR}}$$

suggesting, again, that once a prior for $SR^2_{Y(P).X}$ and a prior for either $R^2_{Y(P,X)}$ or $R^2_{Y(X)}$ are determined, the remaining parameter can be derived for a specified d_{PR}.

For transforming β_1 into $R^2_{Y(P,X)}$ and $R^2_{Y(X)}$, recall that

$$|\beta_1| = \sqrt{\frac{SR^2_{Y(P).X}}{1 - R^2_{P(X)}}} \cdot \frac{\sigma_Y}{\sigma_P}$$

Then rearranging the equation results in

$$SR^2_{Y(P).X} = R^2_{Y(P,X)} - R^2_{Y(X)} = \beta_1^2 \cdot \left(\frac{\sigma_P}{\sigma_Y}\right)^2 \cdot \left(1 - R^2_{P(X)}\right)$$

which is now expressed in terms of $SR^2_{Y(P).X}$. We already know how to determine $R^2_{Y(P,X)}$ and $R^2_{Y(X)}$ with a prior for $SR^2_{Y(P).X}$ and a prior for either $R^2_{Y(P,X)}$ or $R^2_{Y(X)}$. The major difference from the transformation from $R^2_{Y(P).X}$ or $SR^2_{Y(P).X}$ into two squared multiple correlation coefficients $R^2_{Y(P,X)}$ and $R^2_{Y(X)}$ is that the transformation from β_1 requires priors for three additional parameters, namely, $R^2_{P(X)}$, σ_Y, and σ_P. Then $R^2_{Y(P,X)}$ and $R^2_{Y(X)}$ are determined as

$$R^2_{Y(P,X)} = R^2_{Y(X)} + d^2_\beta \cdot \left(\frac{\sigma_P}{\sigma_Y}\right)^2 \cdot \left(1 - R^2_{P(X)}\right)$$

and

$$R^{2*}_{Y(X)} = R^2_{Y(P,X)} - d^2_\beta \cdot \left(\frac{\sigma_P}{\sigma_Y}\right)^2 \cdot \left(1 - R^2_{P(X)}\right)$$

based on priors for $R^2_{P(X)}$, σ_Y, and σ_P for a specified d_β.

Finally, we demonstrate how to transform $\beta_{s,1}$ into $R^2_{Y(P,X)}$ and $R^2_{Y(X)}$. Recall that $\beta_{s,1}$ can be expressed in terms of β_1 as

$$\beta_1 = \beta_{s,1} \cdot \frac{\sigma_Y}{\sigma_P}$$

We already know how to transform β_1 into $R^2_{Y(P,X)}$ and $R^2_{Y(X)}$ with a prior for $SR^2_{Y(P).X}$ and a prior for either $R^2_{Y(P,X)}$ or $R^2_{Y(X)}$. For a specified value d_{β_s}, we obtain $R^2_{Y(P,X)}$ and $R^2_{Y(X)}$ by substituting $\beta_{s,1} \cdot \frac{\sigma_Y}{\sigma_P}$ for β_1 as

$$R^2_{Y(P,X)} = R^2_{Y(X)} + d^2_{\beta_s} \cdot \left(1 - R^2_{P(X)}\right)$$

and

$$R^2_{Y(X)} = R^2_{Y(P,X)} - d^2_{\beta_s} \cdot \left(1 - R^2_{P(X)}\right)$$

While the transformation of $\beta_{s,1}$ into $R^2_{Y(P,X)}$ and $R^2_{Y(X)}$ does not require priors for σ_Y or σ_P as it does for β_1, it requires a prior for $R^2_{P(X)}$, in addition to a prior for either $R^2_{Y(X)}$ or $R^2_{Y(P,X)}$.

In summary, we see that transformation of different parameter specifications into $R^2_{Y(P,X)}$ and $R^2_{Y(X)}$ requires additional parameter specifications. The transformation from either $R^2_{Y(P).X}$ or $SR^2_{Y(P).X}$ is rather simple and straightforward once a prior for either $R^2_{Y(P,X)}$ or $R^2_{Y(X)}$ is determined. On the other hand, the transformation from either β_1 or $\beta_{s,1}$ requires priors for additional parameters, namely, $R^2_{P(X)}$ for $\beta_{s,1}$, and σ_Y, and σ_P in addition to $R^2_{P(X)}$ for β_1. This suggests that the transformation of β_1 requires the largest number of priors for parameters among the transformations of parameters discussed here.

3.3.1.1.9 Sample Size Estimation Using `powerreg`

We are prepared to use the STATA command `powerreg` with our understanding of how to transform $SR^2_{Y(P).X}$, $R^2_{Y(P).X}$, β_1, and $\beta_{s,1}$ into $R^2_{Y(P,X)}$ and $R^2_{Y(X)}$. In this section we demonstrate through numerical examples how to perform sample size estimation in STATA using `powerreg` for the multiple linear regression model of

$$Y = \beta_0 + \beta_1 \cdot P + \beta_2 \cdot X + \varepsilon$$

with hypotheses framed in the regression coefficient β_1 as

$$H_0 : \beta_1 = 0|\beta_2$$

$$H_a : \beta_1 \geq \ d_\beta|\beta_2 \text{ or } \beta_1 \leq \ -d_\beta|\beta_2$$

We focus here on hypotheses framed in β_1 mainly for two reasons. First, the parameter β_1 is the most complicated among the parameters we have reviewed and it requires the largest number of priors for parameters for its transformation into the two multiple correlation coefficients, $R^2_{Y(P,X)}$ and $R^2_{Y(X)}$. Once the procedure of transforming β_1 is mastered, the transformation of the other parameters, namely $R^2_{Y(P).X}$, $SR^2_{Y(P).X}$ and $\beta_{s,1}$, naturally follows. Second, hypotheses are often framed and quantified in terms of β_1 by researchers, reflecting that it is the most intuitive parameter for inference as it represents a change in Y associated with a one-unit change in P. It therefore naturally fits our research question in program evaluation, such as "what is the difference in the outcome between participant and comparison populations?" and "what is the level of change in the outcome for different levels of P?" Additionally, a prior for β_1 may be found in publications because regression coefficients are routinely reported whereas other parameters are less commonly reported. For these two reasons we demonstrate examples for determining $R^2_{Y(P,X)}$ and $R^2_{Y(X)}$ from β_1 and performing sample size estimation in STATA using `powerreg`. However, it should be noted that hypotheses framed in $R^2_{Y(P).X}$, $SR_{Y(P).X}$ or $\beta_{s,1}$ can be reframed in the same manner by applying the corresponding transformation formulas presented above.

Suppose that our aim is to estimate sample size to perform hypothesis testing on β_1 when $d_\beta = 0.5$ with a significance level $\alpha = 0.05$ (two-sided) and power $1 - \beta = 0.8$. How do we translate the research question into the two squared multiple correlation coefficients $R^2_{Y(X)}$ and $R^2_{Y(P,X)}$? As noted above, when hypotheses are framed and quantified in β_1, we need to determine priors for $R^2_{P(X)}$, σ_Y, and σ_P, in addition to either $R^2_{Y(X)}$, or $R^2_{Y(P,X)}$, to obtain both $R^2_{Y(X)}$ and $R^2_{Y(P,X)}$. Suppose we have prior knowledge (which makes us extremely lucky; we will discuss this point later) that $\sigma_Y = 12$, $\sigma_P = 6$, $R^2_{P(X)} = 0.2$, and $R^2_{Y(P,X)} = 0.3$. Then the remaining parameter specification we need is $R^2_{Y(X)}$, which can be determined from the available information as

$$R^2_{Y(X)} = R^2_{Y(P,X)} - d^2_\beta \cdot \left(\frac{\sigma_P}{\sigma_Y}\right)^2 \cdot \left(1 - R^2_{P(X)}\right)$$

$$= 0.3 - (0.5)^2 \cdot \left(\frac{6}{12}\right)^2 \cdot (1 - 0.2) = 0.25$$

implying that $SR^2_{Y(P).X} = R^2_{Y(P,X)} - R^2_{Y(X)} = 0.05$

To estimate sample size for the parameter specifications of $R^2_{Y(P,X)} = 0.3$, $R^2_{Y(X)} = 0.25$, $k = 2$ (the total number of regressors in the model, which is 2 in our example given the regressors P and X), $q = 1$ (the number of regressors tested simultaneously, which is 1 in our example given only P is tested), $1 - \beta = 0.8$, and $\alpha = 0.05$ (two-sided), we feed the powerreg command as:

```
    powerreg, r2f(0.3) r2r(0.25) nvar(2) ntest(1) power(0.8)
alpha(0.05)
```

which produces STATA Output 3.5.

STATA Output 3.5

```
Linear regression power analysis
alpha=.05  nvar=2  ntest=1
R2-full=.3  R2-reduced=.25  R2-change=0.0500

nominal     actual
 power       power          n
 0.8000     0.8005        112
```

Let us examine the output. Below the first line “Linear regression power analysis,” we see parameter values that we specified: the statistical significance level in alpha = .05, k in nvar = 2, q in ntest = 1, $R^2_{Y(P,X)}$ in R2-full = .3, $R^2_{Y(X)}$ in R2-reduced = .25, and $R^2_{Y(P,X)} - R^2_{Y(X)}$ in R2-change = .05 (i.e., 0.3 - 0.25). We refer to the last output n to obtain the estimated sample size, which is 112 for the specified values in this example.

3.3.1.1.10 Effects of Parameter Specifications on Sample Size

We have discussed that statistical hypotheses framed and quantified in β_1 require specifications of additional parameters, namely $R^2_{P(X)}$, σ_Y, and σ_P, in addition to either $R^2_{Y(P,X)}$ or $R^2_{Y(X)}$. Our natural question is: What is the effect of these parameter specifications on estimated sample size? In this section we review each parameter specification, namely d_β, $R^2_{P(X)}$, σ_Y, and σ_P, and its effect on estimated sample size by increasing each prior while holding all other parameter specifications constant. We use the specifications we adopted in the previous section as a reference against which we compare estimated sample sizes based on a change in the parameter specifications. The reference value for each parameter specification and estimated sample size are summarized in Table 3.4 below:

Table 3.4. Reference values

Parameter	**Specification**
$R^2_{Y(P,X)}$	0.3
d_β	0.5
$R^2_{P(X)}$	0.2
σ_P	6
σ_Y	12
k	2
q	1
α (two-sided)	0.05
$1-\beta$	0.80
n	112

First, we examine the relationship between the size of d_β and estimated sample size. Using our running example, suppose that d_β is now 0.6 instead of 0.5, while other parameter specifications remain unchanged. What is the implication to estimated sample size? We reframe the research question in terms of two squared multiple correlation coefficients, $R^2_{Y(X)}$, and $R^2_{Y(P,X)}$, to examine the effect of d_β. Using the same parameter specifications as in our reference example, the remaining parameter $R^2_{Y(P,X)}$ is determined as

$$R^2_{Y(X)} = R^2_{Y(P,X)} - d^2_\beta \cdot \left(\frac{\sigma_P}{\sigma_Y}\right)^2 \cdot \left(1 - R^2_{P(X)}\right)$$

$$= 0.3 - (0.6)^2 \cdot \left(\frac{6}{12}\right)^2 \cdot (1 - 0.2) = 0.228$$

$R^2_{Y(X)}$ is now 0.228 instead of 0.25 as it was when $d_\beta = 0.5$, resulting in $SR^2_{Y(P).X} = R^2_{Y(P,X)} - R^2_{Y(X)} = 0.072$, which is larger by 0.022 compared to when $d_\beta = 0.5$. Let us examine the implication of the change in d_β to estimated sample size. The sample size for power $(1-\beta)$ of 0.8, $R^2_{Y(P,X)} = 0.3$, $R^2_{Y(X)} = 0.1$, $k = 2$, $q = 1$, and $\alpha = 0.05$ (two-sided) can be estimated with the `powerreg` command as:

```
    powerreg, r2f(0.3) r2r(0.228) nvar(2) ntest(1) power(0.8)
alpha(0.05)
```

which produces STATA Output 3.6. The estimated sample size is 78, which is fewer by 34 compared to when $d_\beta = 0.5$.

STATA Output 3.6

```
Linear regression power analysis
alpha=.05  nvar=2  ntest=1
R2-full=.3  R2-reduced=.228  R2-change=0.0720

nominal     actual
 power       power          n
 0.8000     0.7984         78
```

We now examine the relationship between the size of $R^2_{P(X)}$ and estimated sample size. Recall that $1 - R^2_{P(X)}$ represents the level of "tolerance" of P with X. Suppose that P has a lower tolerance and that $R^2_{P(X)}$ is now 0.4 instead of 0.2, resulting in the tolerance of 0.6 instead of 0.8. To examine the effect of $R^2_{P(X)}$ on estimated sample size, we need to obtain $R^2_{Y(P,X)}$ and $R^2_{Y(X)}$ given $R^2_{P(X)} = 0.4$. Suppose $R^2_{Y(P,X)} = 0.3$ remains the same, as well as the rest of the parameters. Then $R^2_{Y(X)}$ is determined as

$$R^2_{Y(X)} = R^2_{Y(P,X)} - d^2_\beta \cdot \left(\frac{\sigma_P}{\sigma_Y}\right)^2 \cdot \left(1 - R^2_{P(X)}\right)$$

$$= 0.3 - (0.5)^2 \cdot \left(\frac{6}{12}\right)^2 \cdot (1 - 0.4) = 0.2625$$

Recall that in our reference example with $R^2_{P(X)} = 0.2$, $R^2_{Y(X)}$ was 0.25. This suggests that a higher $R^2_{P(X)}$ (and hence a lower tolerance of P) results in a larger value for $R^2_{Y(X)}$, and a smaller $SR^2_{Y(P).X}$ of 0.0375, while $d_\beta = 0.5$ has not changed. Let us examine the implication of the change in $R^2_{P(X)}$ to estimated sample size. The sample size for power $(1 - \beta)$ of 0.8, $R^{2*}_{Y(P,X)} = 0.3$, $R^{2*}_{Y(X)} = 0.2625$, $k = 2$, $q = 1$, and $\alpha = 0.05$ (two-sided) can be estimated with `powerreg` as:

```
    powerreg, r2f(0.3) r2r(0.2625) nvar(2) ntest(1) power(0.8)
alpha(0.05)
```

which produces STATA Output 3.7. The estimated sample size is 148, which is larger by 36 compared to when $R^2_{P(X)} = 0.2$.

We now examine the effect of variation in Y on estimated sample size, focusing on two types of Y: a continuous variable Y_c, and a binary variable Y_b. First, we discuss a continuous variable Y_c.

The standard deviation of Y_c is defined as

$$\sigma_{Y_c} = \sqrt{\frac{\sum_{i=1}^{N}\left(Y_{c,i} - \overline{Y}_c\right)^2}{N}}$$

for a population size of N and $\overline{Y}_c = \frac{\sum_{i=1}^{N} Y_{c,i}}{N}$ where $Y_{c,i}$ denotes Y_c of i^{th} individual for $i = 1, ..., N$. Note that when Y_i are more variable around the average ($\overline{Y}_c$), σ_{Y_c} is increased compared to when $Y_{c,i}$ are distributed tightly around the average for a constant N. The magnitude of σ_{Y_c} depends on the characteristics of Y and its unit of measure. For instance, say, the height of college textbooks is perhaps somewhat standardized (for the convenience of users and publishers) and less variable compared to their thickness, which is determined by the content and perhaps less controlled by publishers. At the same time, the unit of measure matters. Recall that

$$\sigma^2_{a \cdot Y} = a^2 \sigma^2_Y$$

for a constant a. This implies

$$\sqrt{\sigma^2_{a \cdot Y}} = \sigma_{a \cdot Y_c} = a \cdot \sigma_{Y_c}$$

Therefore the standard deviation of, say, the height of college textbooks differ depending on the unit of measurement (inches, centimeters, etc.).[80] In theory, σ_{Y_c} can take any value in $[0, \infty)$.

STATA Output 3.7

```
Linear regression power analysis
alpha=.05  nvar=2  ntest=1
R2-full=.3  R2-reduced=.2625  R2-change=0.0375

nominal     actual
 power       power          n
 0.8000     0.7987        148
```

Now that we understand the nature of σ_{Y_c}, let us examine the effect of σ_{Y_c} on estimated sample size. Using our running example, suppose σ_{Y_c} is now 18 instead of 12, higher by 6. To examine the effect of this change on estimated sample size, we need to specify the remaining parameter value,

[80] One measure for assessing the *relative* distribution of a variable is a coefficient of variation, denoted as cv, which is defined as

$$cv = \frac{\sigma_Y}{\overline{Y}}$$

The coefficient of variation measures the size of standard deviation of Y (σ_Y) relative to the average of Y ($\overline{Y}$). The measure is unit-free and often used to compare standard deviations of different variables measured in different units.

$R^2_{Y(X)}$, while keeping other priors for parameters unchanged. $R^2_{Y(X)}$ is determined as

$$R^2_{Y(X)} = R^2_{Y(P,X)} - d_\beta^2 \cdot \left(\frac{\sigma_P}{\sigma_Y}\right)^2 \cdot \left(1 - R^2_{P(X)}\right)$$

$$= 0.3 - (0.5)^2 \cdot \left(\frac{6}{18}\right)^2 \cdot (1 - 0.2) \approx 0.278$$

implying that $SR^2_{Y(P).X} = R^2_{Y(P,X)} - R^2_{Y(X)} \approx 0.3 - 0.278 = 0.022$, smaller by about 0.028 compared to when $\sigma_{Y_c} = 12$. Let us examine the implication of the change in σ_{Y_c} to estimated sample size. The sample size for power $(1 - \beta)$ of 0.8, $R^2_{Y(P,X)} = 0.3$, $R^2_{Y(X)} = 0.278$, $k = 2$, $q = 1$ and $\alpha = 0.05$ (two-sided) can be estimated with the `powerreg` command as:

```
    powerreg, r2f(0.3) r2r(0.278) nvar(2) ntest(1) power(0.8)
alpha(0.05)
```

which produces STATA Output 3.8. The estimated sample size is 252, which is larger by 140 compared to when $\sigma_{Y_c} = 12$.

STATA Output 3.8

```
Linear regression power analysis
alpha=.05  nvar=2  ntest=1
R2-full=.3  R2-reduced=.278  R2-change=0.0220

nominal     actual
 power       power          n
 0.8000     0.8005        252
```

Next, we discuss a binary variable Y_b. For Y_b, $Y_{b,i}$ (which denotes Y_b of i^{th} individual) takes a value of either 0 or 1. The standard deviation of Y_b is defined as

$$\sigma_{Y_b} = \sqrt{\overline{Y}_b \left(1 - \overline{Y}_b\right)}$$

where $\overline{Y}_b = \frac{\sum_{i=1}^{N} Y_{b,i}}{N}$ for a population size N. Note that $\overline{Y}_b$ is contained in $[0, 1]$, and σ_{Y_b} is maximized at 0.5 when $\overline{Y}_b = 1 - \overline{Y}_b = 0.5$. That is, σ_{Y_b} is contained in $[0, 0.5]$, unlike σ_{Y_c} which in theory can take any value in $[0, \infty)$. To examine its effect on estimated sample size, we cannot use our running example as a reference because $\sigma_{Y_b} = 12$ is larger than the maximum value of 0.5 that σ_{Y_b} can take. We therefore set our reference σ_{Y_b} at 0.4 first. Also we specify $d_\beta = 0.02$, which is more realistic given a binary outcome Y_b and a standard deviation of P at 6.[81] We keep all the

[81] Notice that with $d_\beta = 0.5$, which was used in our running example, it means that a change by two units in P, which is only one-third of its standard deviation, is associated with an expected change by 1.0 in $\overline{Y}_b$. Likewise, a change by three units in P, which is only half of its standard deviation, is associated with an expected change by 1.5 in $\overline{Y}_b$. Therefore $d_\beta = 0.5$ may not be a realistic specification as a statistical model given that $\overline{Y}_b$ is contained in

other parameter specifications unchanged. Let's see what an increase by 0.1 in σ_{Y_b} to 0.5 does on estimated sample size.

The reference values of $\sigma_{Y_b} = 0.4$ and $d_\beta = 0.02$ imply $R^2_{Y(X)} = 0.228$ because

$$R^2_{Y(X)} = R^2_{Y(P,X)} - d^2_\beta \cdot \left(\frac{\sigma_P}{\sigma_Y}\right)^2 \cdot \left(1 - R^2_{P(X)}\right)$$

$$= 0.3 - (0.02)^2 \cdot \left(\frac{6}{0.4}\right)^2 \cdot (1 - 0.2) = 0.228$$

The sample size for the same parameter specifications of $R^2_{Y(P,X)} = 0.3$, $R^2_{Y(X)} = 0.228$, $k = 2$, $q = 1$, $1 - \beta = 0.8$ and $\alpha = 0.05$ is estimated by feeding the `powerreg` command as:

```
    powerreg, r2f(0.3) r2r(0.228) nvar(2) ntest(1) power(0.8)
alpha(0.05)
```

The estimated sample size for our reference of $\sigma_{Y_b} = 0.4$ is 78 as shown in STATA Output 3.9.

STATA Output 3.9

```
Linear regression power analysis
alpha=.05  nvar=2  ntest=1
R2-full=.3  R2-reduced=.228  R2-change=0.0720

nominal     actual
 power       power          n
 0.8000     0.7984         78
```

Now that we have estimated the sample size for our reference case, we are ready to examine the effect of a change in σ_{Y_b}. When σ_{Y_b} is increased from 0.4 to 0.5, $R^2_{Y(X)} \approx 0.254$ because

$$R^2_{Y(X)} = R^2_{Y(P,X)} - d^2_\beta \cdot \left(\frac{\sigma_P}{\sigma_Y}\right)^2 \cdot \left(1 - R^2_{P(X)}\right)$$

$$= 0.3 - (0.02)^2 \cdot \left(\frac{6}{0.5}\right)^2 \cdot (1 - 0.2) \approx 0.254$$

Let us examine the implication of the change in σ_{Y_b} to estimated sample size. The sample size for power $(1 - \beta)$ of 0.8, $R^2_{Y(P,X)} = 0.3$, $R^2_{Y(X)} = 0.254$, $k = 2$, $q = 1$ and $\alpha = 0.05$ can be estimated with the `powerreg` command as:

```
    powerreg, r2f(0.3) r2r(0.254) nvar(2) ntest(1) power(0.8)
alpha(0.05)
```

which produces STATA Output 3.10. The estimated sample size is 120, which is larger by 42

$[0, 1]$.

compared to our reference of 78 when $\sigma_{Y_b} = 0.4$.

STATA Output 3.10

```
Linear regression power analysis
alpha=.05  nvar=2  ntest=1
R2-full=.3  R2-reduced=.254  R2-change=0.0460

nominal     actual
 power       power          n
 0.8000     0.7953        120
```

We now examine the effect of variation in P on estimated sample size. As we did for Y, we discuss two types of P (a continuous variable P_c and a binary variable P_b) in order to understand the effect of variation in P on estimated sample size. First, a continuous variable P_c has a standard deviation defined as

$$\sigma_{P_c} = \sqrt{\frac{\sum_{i=1}^{N}\left(P_{c,i} - \overline{P}_c\right)^2}{N}}$$

for a population size of N and $\overline{P}_c = \frac{\sum_{i=1}^{N} P_{c,i}}{N}$ where $P_{c,i}$ denotes P_c of i^{th} individual for $i = 1, ..., N$. Again, as it is the case for σ_{Y_c}, σ_{P_c} is increased when P_i are more variable around the average ($\overline{P}_c$) compared to when $P_{c,i}$ are distributed tightly around the average for a constant N. In theory, σ_{P_c} can take any value in $[0, \infty)$. Using our running example, suppose σ_{P_c} is now 8 instead of 6, larger by 2. To examine the effect of this change on estimated sample size, we need to specify the remaining parameter value, $R^2_{Y(X)}$, while keeping other parameter specifications unchanged. $R^2_{Y(X)}$ is determined as

$$\begin{aligned} R^2_{Y(X)} &= R^2_{Y(P,X)} - d^2_{\beta} \cdot \left(\frac{\sigma_P}{\sigma_Y}\right)^2 \cdot \left(1 - R^2_{P(X)}\right) \\ &= 0.3 - (0.5)^2 \cdot \left(\frac{8}{12}\right)^2 \cdot (1 - 0.2) \approx 0.211 \end{aligned}$$

implying that $SR^2_{Y(P).X} = R^2_{Y(P,X)} - R^2_{Y(X)} \approx 0.3 - 0.211 \approx 0.089$, larger by about 0.039 compared to when $\sigma_{P_c} = 6$.

Let us examine the implication of the change in σ_{P_c} to estimated sample size. The sample size for power of 0.8, $R^2_{Y(P,X)} = 0.3$, $R^2_{Y(X)} = 0.211$, $k = 2$, $q = 1$ and $\alpha = 0.05$ (two-sided) can be estimated with the `powerreg` as:

```
    powerreg, r2f(0.3) r2r(0.211) nvar(2) ntest(1) power(0.8)
alpha(0.05)
```

which produces STATA Output 3.11. The estimated sample size is 63, which is smaller by 59 compared to when $\sigma_{P_c} = 6$.

STATA Output 3.11

```
Linear regression power analysis
alpha=.05  nvar=2  ntest=1
R2-full=.3  R2-reduced=.211  R2-change=0.0890

nominal     actual
 power       power          n
 0.8000     0.7952         63
```

Next, let us discuss a binary variable P_b. For P_b, $P_{b,i}$ (which denotes P_b of i^{th} individual) takes a value of either 0 or 1. The standard deviation of P_b is defined as

$$\sigma_{P_b} = \sqrt{\overline{P}_b\left(1 - \overline{P}_b\right)}$$

where $\overline{P}_b = \frac{\sum_{i=1}^{N} P_{b,i}}{N}$ for a population size N. $\overline{P}_b$ is contained in $[0, 1]$, and σ_{P_b} is maximized at 0.5 when $\overline{P}_b = 1 - \overline{P}_b = 0.5$. For examining the effect of a change in σ_{P_b} on estimated sample size, we will again tweak our running example, as we did for a binary outcome Y_b, to make it a more statistically realistic model given the distribution of P_b. First we will set our reference value of σ_{P_b} to 0.4 because $\sigma_{P_b} = 6$, which was used in our running example, is larger than the maximum value of 0.5 that σ_{P_b} can take. Also we specify $d_\beta = 6$ instead of $d_\beta = 0.5$.[82] We keep all the other parameter specifications unchanged.

In order to examine the effect of an increase in σ_{P_b} from 0.4 to 0.5, we first need to evaluate the estimated sample size for the reference values of $\sigma_{P_b} = 0.4$ and $d_\beta = 6$, which result in $R^2_{Y(X)} = 0.268$ because

$$R^2_{Y(X)} = R^2_{Y(P,X)} - d_\beta^2 \cdot \left(\frac{\sigma_P}{\sigma_Y}\right)^2 \cdot \left(1 - R^2_{P(X)}\right)$$

$$= 0.3 - (6)^2 \cdot \left(\frac{0.4}{12}\right)^2 \cdot (1 - 0.2) = 0.268$$

The sample size for $R^2_{Y(P,X)} = 0.3$, $R^2_{Y(X)} = 0.268$, $n = 150$, $k = 2$, $q = 1$ and $\alpha = 0.05$ with power of 0.8 is estimated by feeding the `powerreg` command as:

```
    powerreg, r2f(0.3) r2r(0.268) nvar(2) ntest(1) power(0.8)
alpha(0.05)
```

[82] The specification of $d_\beta = 0.5$, which we used in our running example, posed no model conflict and we could have kept it unchanged. However, the specification means that a change from 0 to 1 in P is associated with a change by 0.5 in $\overline{Y}$, which is only 4% of the standard deviation of Y. As a result, the estimated statistical power would be very small and the estimated sample size would be large. We found it hard to clearly demonstrate the effect of a change in σ_{P_b} on estimated statistical power or sample size. Therefore, we changed the specification of d_β from 0.5 to 6 for the purpose of demonstration.

The estimated sample size for our reference of $\sigma_{P_b} = 0.4$ is 172 as shown in STATA Output 3.12.

STATA Output 3.12

```
Linear regression power analysis
alpha=.05  nvar=2  ntest=1
R2-full=.3  R2-reduced=.268  R2-change=0.0320

nominal     actual
 power       power          n
 0.8000     0.7962        172
```

Now that we have estimated the sample size for our reference case, we are ready to examine the effect of a change in σ_{P_b}. When σ_{P_b} is increased from 0.4 to 0.5, $R^2_{Y(X)} = 0.25$ because

$$R^2_{Y(X)} = R^2_{Y(P,X)} - d^2_\beta \cdot \left(\frac{\sigma_P}{\sigma_Y}\right)^2 \cdot \left(1 - R^{2*}_{P(X)}\right)$$

$$= 0.3 - (6)^2 \cdot \left(\frac{0.5}{12}\right)^2 \cdot (1 - 0.2) = 0.25$$

Let us examine the implication of the change in σ_{P_b} to the estimated sample size. The sample size for power of 0.8, $R^2_{Y(P,X)} = 0.3$, $R^2_{Y(X)} = 0.25$, $k = 2$, $q = 1$ and $\alpha = 0.05$ (two-sided) can be estimated with the `powerreg` command as:

```
    powerreg, r2f(0.3) r2r(0.25) nvar(2) ntest(1) power(0.8)
alpha(0.05)
```

which produces STATA Output 3.13. The estimated sample size is 112, which is smaller by 60 compared to our reference of 172 when $\sigma_{P_b} = 0.4$.

STATA Output 3.13

```
Linear regression power analysis
alpha=.05  nvar=2  ntest=1
R2-full=.3  R2-reduced=.25  R2-change=0.0500

nominal     actual
 power       power          n
 0.8000     0.8005        112
```

We can now derive the general rule of how parameter specifications influence estimated sample size. While we consistently increased the specified parameter value in these examples, we see that some increased the estimated sample size, while the others worked in the opposite direction. What is the inference we can draw here? To understand the general rule of how parameter specifications influence estimated sample size, it is intuitive to think in terms of the difference between $R^2_{Y(P,X)}$ and $R^2_{Y(X)}$, or equivalently $SR^2_{Y(P).X}$. Recall that

$$R^2_{Y(P,X)} - R^2_{Y(X)} = SR^2_{Y(P).X}$$

and

$$SR^2_{Y(P).X} = d^2_\beta \cdot \left(\frac{\sigma_P}{\sigma_Y}\right)^2 \cdot \left(1 - R^2_{P(X)}\right)$$

Also recall that our alternative hypotheses framed in terms of β_1 as

$$H_a : \beta_1 \geq \ d_\beta|\beta_2 \text{ or } \beta_1 \leq \ -d_\beta|\beta_2$$

are statistically equivalent to

$$H_a : SR^2_{Y(P).X} \geq d_{SR}$$

This implies that the larger is d_{SR}, the further away is d_{SR} from 0 (the null value). As we reviewed in the first section of Chapter 3, this results in a smaller sample size (for a given statistical power). Therefore, any parameter specifications that result in an increase in $SR^2_{Y(P).X}$ result in a smaller estimated sample size, while those that result in a decrease in $SR^2_{Y(P).X}$ result in a larger estimated sample size. This is why we have seen that an increase in parameters that comprise the numerator of $SR^2_{Y(P).X}$ (i.e., d_β, σ_P, and $1 - R^2_{P(X)}$) result in a decrease in the estimated sample size, while an increase in σ_Y, which comprises the denominator of $SR^2_{Y(P).X}$, worked in the opposite way. However, note that the sample size formula is based on a nonlinear function. Therefore, predicting the sample size implications of a change in one or more of these parameters, even holding the others, can be tricky. One should always actually perform the calculations to see what the net impact of the changes will be.

Additionally, recall from the first section of this chapter that estimated sample size and power have an inverse relationship, which implies that a change in a parameter specification works in an opposite direction between estimated sample size and power, holding other parameter specifications constant. That is, when an increase in a parameter specification results in an increase in estimated power (for a given sample size), it results in a decrease in estimated sample size (and vice versa). This is intuitive given that a larger sample size is required for a larger statistical power, holding other parameter specifications constant.

Table 3.5 summarizes the direction of the effect of each parameter specification on estimated sample size and power. For the simplicity of the presentation, we show only the effect of an increase in each parameter specification (while holding other parameter specifications constant) on estimated power and sample size. While we do not show the effects of decrease in these parameter specifications, they can be deduced by reversing the direction in Table 3.5.

Table 3.5. Effects of parameter specifications

Increase in	Estimated Power	Estimated Sample Size
d_β	↑	↓
σ_P	↑	↓
σ_Y	↓	↑
$R^2_{P(X)}$	↓	↑

3.3.1.1.11 How to Find Priors for σ_Y, σ_P, and $R^2_{P(X)}$

In the previous section we estimated sample size based on hypothetical examples of parameter specifications. We demonstrated that parameter specifications influence estimated sample size, which implies that it is crucial to adopt accurate priors for σ_Y, σ_P, and $R^2_{P(X)}$. The natural question that follows is: where and how do we find the priors? In practice, these values may not be readily available to us and may need to be "guesstimated" from past experience, literature, and data sources that provide these priors or approximations.

One possible approach to obtaining σ_Y and σ_P is to approximate them based on assumptions about their distributions. We will use P as an example but the same method can be applied to Y. Suppose P takes a limited range of values, say 10 levels, and P_j denotes the value of P at j^{th} level for $j = 1, ..., 10$. Suppose that the proportion of individuals exposed to each level of P_j is W_j such that $\sum_{j=1}^{10} W_j = 1$. For instance, if 25% of individuals in a study are expected to be exposed to P_1, then $W_1 = 0.25$. Then the standard deviation of P (σ_P) can be derived as

$$\sigma_P = \sqrt{\sum_{j=1}^{10} W_j \cdot \left(P_j - \overline{P}\right)^2}$$

where $\overline{P}$ can be obtained as $\overline{P} = \sum_j W_j \cdot P_j$. Notice that when P is binary (takes a value of either 0 or 1) as it is in a study with two groups, that is, participant and comparison populations, the derivation reduces to

$$\sigma_P = \sqrt{W_1 \cdot W_0}$$

where W_0 and W_1 are the proportion of program participants and non-participants, respectively, such that $W_0 + W_1 = 1$.[83]

This method is useful because the information we need is the range of levels that the variable

[83] The proof is provided as follows. Let $P_0 = 0$ and $P_1 = 1$. Notice that $\overline{P} = \sum_j W_j \cdot P_j = W_0 \cdot 0 + W_1 \cdot 1 = W_1$. Then σ_P can be rewritten as

$$\sigma_P = \sqrt{\sum_j W_j \cdot \left(P_j - \overline{P}\right)^2} = \sqrt{W_0 \cdot \left(0 - \overline{P}\right)^2 + W_1 \cdot \left(1 - \overline{P}\right)^2}$$

$$= \sqrt{W_0 \cdot (0 - W_1)^2 + W_1 \cdot (1 - W_1)^2} = \sqrt{W_0 \cdot W_1^2 + W_1 \cdot W_0^2}$$
$$= \sqrt{W_1 \cdot W_0 \cdot (W_0 + W_1)} = \sqrt{W_1 \cdot W_0 \cdot (1)} = \sqrt{W_1 \cdot W_0}$$

thus establishing the proof.

can take and the proportion of individuals that belong to each value of the variable. When the variable is under the control of researchers, this may be assumed quite accurately from the study protocol. For instance, in a dose-response type of analysis, a researcher may know the level of dose (P_j) that will be administered and the proportion of participants that will receive each dose level (W_j). Then the standard deviation can be estimated based on the information.

When accurate priors for σ_Y, σ_P, and $R^2_{P(X)}$ cannot be obtained, it is recommended to adopt "conservative" specifications. We have demonstrated that larger σ_Y and $R^2_{P(X)}$, and smaller σ_P result in a smaller estimated statistical power and a larger estimated sample size. Then it is helpful to perform power and sample size estimations on a range of parameter specifications that include conservative values because it will provide ideas as to the sensitivity of the estimated power and sample size, as well as their conservative estimations. One useful approach when Y is binary is to adopt $\sigma_Y = 0.5$ — the maximum value — to obtain the minimum estimated power and the maximum sample size given a set of specifications of other parameters.

3.3.1.1.12 RCTs and Regression Modelling

We briefly digress to discuss a seemingly curious popular application of regression analysis. It is not at all uncommon to see regression analyses applied to samples generated from RCTs (for example, random samples of participants and non-participants from an RCT). Indeed, a model more or less along the lines of

$$Y_i = \beta_0 + \beta_1 \cdot P_i + \beta_2 \cdot X_i + \varepsilon_i$$

is often applied in the setting of RCTs.

This is a common practice even in cases where evaluators claim that the sample is experimental in the sense of Lance, Guilkey, et al. (2014): the variation in program exposure across the population of participants and non-participants, and hence across random samples from them, is random.

This is a rather curious practice since perhaps the central selling point of an RCT is that it allows us to ignore variation in observables (such as X) and whatever unobservables are relegated to ε for the purpose of determining the impact (i.e., causal influence) of P on the outcome Y. At first glance, the use of the aforementioned regression model in the setting of RCTs might seem unnecessary in a sense that it controls for X.

Indeed, from the standpoint of recovering an *unbiased* estimated of β_1, and hence program impact, it generally *is* unnecessary to estimate a regression model that controls for X. The argument for applying such a model to samples generated by RCTs instead rests on precision. The argument can (and will) be formally presented, but basically on a conceptual level it boils down to this: By controlling for X we remove some variation in Y, in the process lowering sampling variation in $\hat{\beta}_1$ by focusing on the remaining, less substantial, variation in Y after controlling for X.

To see this, let's begin with a rather simple linear regression model in this example defined as

$$\text{Model 1} : Y = \alpha_0 + \alpha_1 \cdot P + \varepsilon_1$$

where we suppress the unit of observation index variable i since it adds little but notational complexity to the discussion to follow in the next few paragraphs. The variance of $\hat{\alpha}_1$ is

$$var\left(\hat{\alpha}_1\right) = \frac{\sigma_1^2}{\left(\sum_{i=1}^{n}\left(P_i - \bar{P}\right)^2\right)}$$

where σ_1^2 is the variance around the regression line with its estimate

$$\hat{\sigma}_1^2 = \frac{SS_{RESID[Y(P)]}}{n-2}$$

This is the fairly standard single regressor setup.

Our multiple linear regression model is

$$\text{Model 2} : Y = \beta_0 + \beta_1 \cdot P + \beta_2 \cdot X + \varepsilon_2$$

where we introduce the subscript 2 to the error term to differentiate it from the error term from the simple single regressor (P) model of the preceding paragraph. The variance of $\hat{\beta}_1$ is

$$var\left(\hat{\beta}_1\right) = \frac{\sigma_2^2}{\left(\sum_{i=1}^{n}\left(P_i - \bar{P}\right)^2\right)}$$

where σ_2^2 is the variance about the regression line with its estimate

$$\hat{\sigma}_2^2 = \frac{SS_{RESID[Y(P,X)]}}{n-3}$$

Note that $R_{P(X)}^2 = 0$ is assumed given we are talking about an RCT so that $1 - R_{P(X)}^2 = 1$.

$\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ differ across the two models in two respects. First, the two involve the two different sum of squared residuals $SS_{RESID[Y(P)]}$ (in Model 1) and $SS_{RESID[Y(P,X)]}$ (in Model 2). Second, they have different denominators (which are actually degrees of freedom) of $n-2$ (in Model 1) and $n-3$ (in Model 2). We can compare $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ along these lines of difference.

First,

$$SS_{RESID[Y(P,X)]} < SS_{RESID[Y(P)]}$$

when X is linearly associated with Y after accounting for P. Second,

$$n - 3 < n - 2$$

Therefore, we have two factors that affect the relationship between $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ in opposite ways. When the reduction in SS_{RESID} is substantial compared to the loss in the degrees of freedom

$$var\left(\hat{\beta}_1\right) < var\left(\hat{\alpha}_1\right)$$

Specifically, we gain precision by including X in a regression model when[84]

$$SS_{RESID[Y(P,X)]} < (n-3) \cdot \frac{SS_{RESID[Y(P)]}}{n-2}$$

In general, and particularly if n is large, we would expect the numerator effect to overwhelm the denominator effect. In Figure 3.17 we graph

$$\frac{n-3}{n-2}$$

against the sample size n. As the figure makes plain, the ratio essentially very rapidly converges to 1. At $n = 22$ it is .95, climbing to 0.979166667 at $n = 50$ before reaching 0.989795918 by $n = 100$. X would have to have very little linear explanatory power for Y for

$$SS_{RESID[Y(P,X)]} < SS_{RESID[Y(P)]}$$

[84] More generally, we gain precision by including control variables $\vec{x}$ in a regression model when $SS_{RESID[Y(P,\vec{X})]} < (n - k_2 - 1) \cdot \frac{SS_{RESID[Y(P)]}}{n-k_1-1}$ where k_1 and k_2 are the total number of regressors in the two models, respectively.

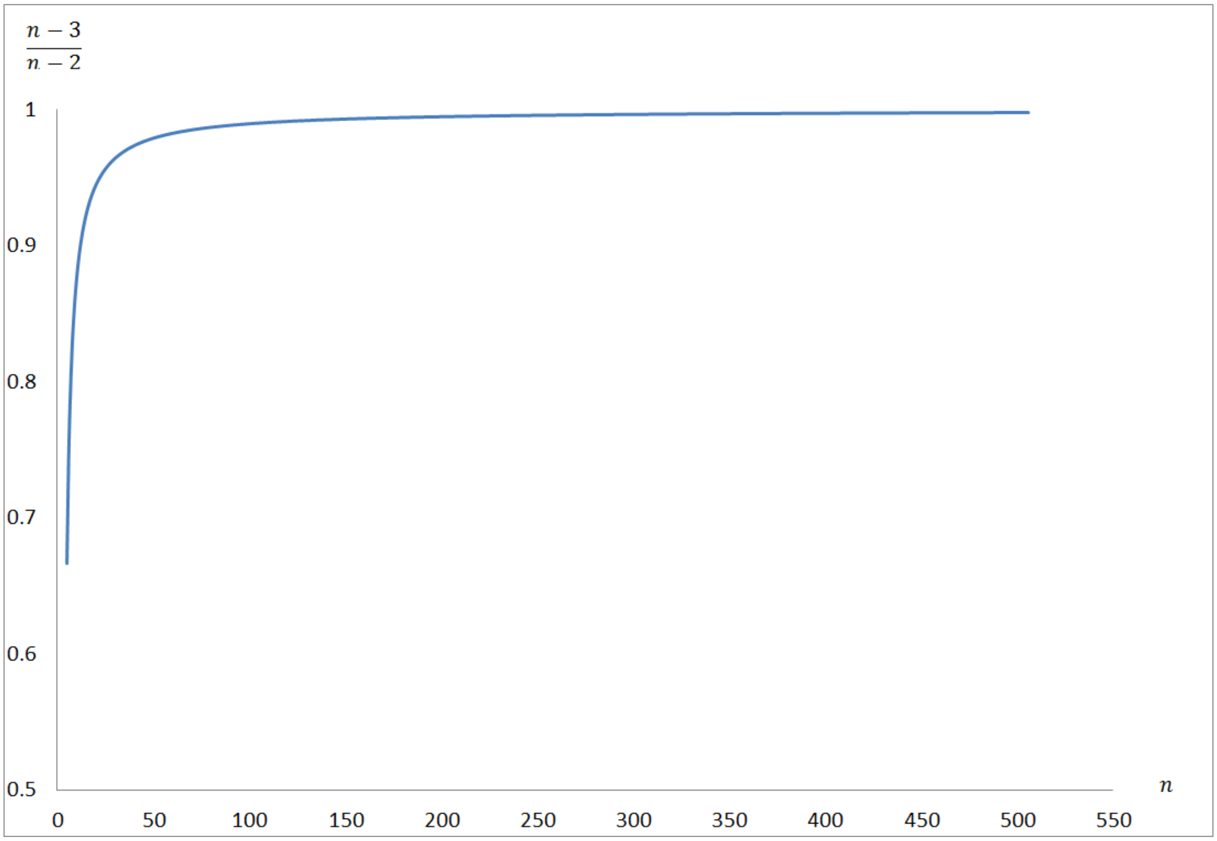


Figure 3.17. The ratio of degrees of freedom as n increases

not to overwhelm

$$\frac{n-3}{n-2}$$

by around $n = 100$.

The point of all of this is that in general one will probably do better (from a sampling variance standpoint) by estimating program impact through the model

$$Y_i = \beta_0 + \beta_1 \cdot P_i + \beta_2 \cdot X_i + \varepsilon_i$$

as opposed to simple regression of Y on P alone or simple comparison of average outcomes between the participant and non-participant samples.

There is thus an argument for thinking about sample size estimation for an RCT from a regression standpoint. To begin with, returning to our original motivating potential outcomes model, we can easily motivate a regression model for RCTs. Once again, we have the potential outcome equations

$$Y_i^0 = \beta_0 + \beta_2 \cdot X_i + \varepsilon_i$$

and

$$Y_i^1 = \beta_0 + \beta_1 + \beta_2 \cdot X_i + \varepsilon_i$$

We also specify the cost of participation as

$$C_i = \gamma_0 + \gamma_1 \cdot X_i + \varepsilon_i^C$$

Thus, the cost of participation now depends on the observed variable as well as an unobserved characteristic represented by ε_i^C. We adopt the earlier condition for determining participation status. Specifically, the individual participates (i.e., $P_i = 1$) if

$$Y_i^1 - Y_i^0 - C_i \geq 0$$

or, inserting the functions we have proposed,

$$\beta_0 + \beta_1 + \beta_2 \cdot X_i + \varepsilon_i - \beta_0 - \beta_2 \cdot X_i - \varepsilon_i - \gamma_0 - \gamma_1 \cdot X_i - \varepsilon_i^C \geq 0$$

or, removing terms,

$$\beta_1 - \gamma_0 - \gamma_1 \cdot X_i - \varepsilon_i^C \geq 0$$

Rearranging, we have

$$-\gamma_1 \cdot X_i - \varepsilon_i^C \geq -\beta_1 + \gamma_0$$

To motivate an RCT, we assume that $\gamma_1 = 0$ and that the designer of the RCT assign each participant in the RCT a value for ε^C of either minus infinity ($-\infty$),[85] in which case they are

[85] Or some suitably large negative number.

program participants, or plus infinity,[86] in which case they are program non-participants/controls. Observed Y is then still just

$$\begin{aligned} Y_i &= P_i \cdot Y_i^1 + (1 - P_i)\, Y_i^0 \\ &= P_i \cdot (\beta_0 + \beta_1 + \beta_2 \cdot X_i + \varepsilon_i) + (1 - P_i) \cdot (\beta_0 + \beta_2 \cdot X_i + \varepsilon_i) \\ &= \beta_0 + \beta_1 \cdot P_i + \beta_2 \cdot X_i + \varepsilon_i \end{aligned}$$

We are thus left with the fairly straightforward setup from the beginning of the linear regression discussion, with the exception that program participation is wholly randomly determined through assignment of ε^C by the designers of the RCT. As we have seen, there is likely to be a benefit from pursuing such a regression approach in terms of the precision of impact estimates, even if the RCT is successful.

3.3.1.2 Logistic Regression

We now turn our attention to dichotomous or binary outcomes, which take only two values. Examples of such outcomes include the status of a condition, for instance having experienced a certain disease, having an income under the poverty threshold, etc. for which the response is coded in two categories, typically "yes" and "no." Then a research question is often framed in terms of the probability of the event of interest. With a binary outcome Y, a typical approach is to define

$$Y_i = 0 \text{ if the event is not observed for } i^{th} \text{ individual}$$

$$Y_i = 1 \text{ if the event is observed for } i^{th} \text{ individual}$$

for $i = 1, ..., N$ of a population size of N. This suggests that $\overline{Y} = \frac{\sum_{i=1}^{N} Y_i}{N}$ is contained in [0,1] for a binary outcome Y. Note that this is a major difference from OLS, which models a continuous variable Y. $E\left[Y|\vec{X}\right]$, the estimated conditional mean of Y given a set of regressors $\vec{X}$, in OLS regression is not limited to any specific range of values and can take a value in $(-\infty, +\infty)$ in theory. However, when the outcome variable Y is dichotomous, it may be intuitive for framing research questions to adopt a statistical approach to modeling the conditional mean of Y as a conditional probability $\Pr\left(Y = 1|\vec{X}\right)$ contained within the range of a probability (i.e., between 0 and 1). Logistic regression is one such regression method that ensures that the conditional mean of Y falls between 0 and 1 while allowing for modeling of the relationship of Y with multiple regressors.[87]

We work with the same variables as we did in the section on OLS: an outcome variable Y, a program variable P, and a control variable X. The main difference from the example in OLS is that Y is now dichotomous, that is, it takes only two values. The relationship of Y with the two

[86] Or, similarly, some suitable large positive value.

[87] For a comprehensive review of logistic regression, see Hosmer, Lemeshow, & Sturdivant, 2013.

regressors P and X in the logistic regression framework can be defined as

$$\Pr(Y=1|P,X)=\pi_{(P,X)}=\frac{e^{[\beta_0+\beta_1\cdot P+\beta_2\cdot X]}}{1+e^{[\beta_0+\beta_1\cdot P+\beta_2\cdot X]}}=\frac{1}{1+e^{[-(\beta_0+\beta_1\cdot P+\beta_2\cdot X)]}}$$

where $\pi_{(P,X)}$ denotes the probability of the event Y, that is, $\Pr(Y=1|P,X)$, and $e^{[\cdot]}$ denotes the exponential function. The right side of the formula is expressed in terms of the logistic function $f(z)$ defined as

$$f(z)=\frac{e^z}{1+e^z}=\frac{1}{1+e^{-z}}$$

where in our example

$$z=\beta_0+\beta_1\cdot P+\beta_2\cdot X$$

Notice that $f(z)\to 1$ as $z\to\infty$ and $f(z)\to 0$ as $z\to-\infty$, that is, the values of $f(z)$ are contained between 0 and 1.

The model $\Pr(Y=1|P,X)=\frac{1}{1+e^{[-(\beta_0+\beta_1\cdot P+\beta_2\cdot X)]}}$ can be rewritten in terms of the logit form as

$$logit\,[\Pr(Y=1|P,X)]=\ln\left[\frac{\pi_{(P,X)}}{1-\pi_{(P,X)}}\right]=\beta_0+\beta_1\cdot P+\beta_2\cdot X$$

where $\ln[\cdot]$ denotes the natural logarithm of an expression.

As was the case in OLS regression, the regression coefficients β_1 and β_2 quantify the relationship between the outcome variable Y and the regressors P and X, respectively. The main difference in modeling between OLS and logistic regression is that β_1 and β_2 are specified in terms of a parameter called "odds ratio" in logistic regression whereas in OLS regression they represent a change in Y associated with a one-unit change in the corresponding regressor. To discuss the specification of logistic regression, we will first review the concept of odds and odds ratio.

Odds is defined as the ratio of the probability that the event of interest Y will occur ($\Pr[Y=1]$) over the probability that the event of interest will not occur ($\Pr[Y=0]=1-\Pr[Y=1]$). Odds is defined as

$$odds_Y=\frac{\Pr(Y=1)}{1-\Pr(Y=1)}$$

Let $odds_{Y(P=1)}$ and $odds_{Y(P=0)}$ denote the odds of event Y observed in two groups: one group with $P=1$ and the other group with $P=0$. Then an odds ratio (OR) is a ratio of two odds of event Y derived from the two groups:

$$OR_{Y(P)}=\frac{odds_{Y(P=1)}}{odds_{Y(P=0)}}=\left[\frac{\Pr(Y=1|P=1)}{1-\Pr(Y=1|P=1)}\right]\Big/\left[\frac{\Pr(Y=1|P=0)}{1-\Pr(Y=1|P=0)}\right]$$

where the subscript $Y(P)$ denotes the event of interest Y with respect to P. The odds ratio describes the size of odds of the first group relative to the odds of the second group.[88] An odds ratio of one

[88] Note that the numerator and the denominator can be switched to estimate the odds ratio in terms of the odds of Group 2 relative to that of Group 1, in other words, $\frac{odds_{Y(P=0)}}{odds_{Y(P=1)}}$

therefore indicates that the odds for the two groups are equivalent — there is no difference in the odds between the two groups.

The logit form of the logistic regression model indicates that the relationships between the odds ratio of the dichotomous outcome variable Y and the regressors P and X are quantified by the regression coefficients β_1 and β_2. Suppose that the program participation status P takes dichotomous values: $P_i = 1$ if i^{th} individual participates in the program of interest, and $P_i = 0$ otherwise. Then the odds ratio of the outcome variable Y for program participation P, holding X constant, is given by

$$OR_{Y(P).X} = \frac{odds_{Y(P=1,X)}}{odds_{Y(P=0,X)}} = \frac{e^{(\beta_0+\beta_1\cdot 1+\beta_2\cdot X)}}{e^{(\beta_0+\beta_1\cdot 0+\beta_2\cdot X)}}$$

$$= e^{[(\beta_0-\beta_0)+\beta_1+(\beta_2\cdot X-\beta_2\cdot X)]} = e^{\beta_1}$$

While the example pertains to a dichotomous status of program participation ($P = 0, 1$), the model can be applied to a model with continuous regressors. For instance, we may be interested in performing dose-response analysis in which P takes many values representing doses of a regime instead of dichotomous values. In such cases, we can apply the logistic regression to assess the relationship between Y and a continuous regressor P and a control variable X as

$$\Pr(Y = 1|P, X) = \frac{1}{1 + e^{[-(\beta_0+\beta_1\cdot P+\beta_2\cdot X)]}}$$

Then e^{β_1} — the exponentiated coefficient for P — gives an odds ratio for comparing groups that differ by one unit in P. That is, e^{β_1} represents an odds ratio related to a one-unit change in P. A more general form of the odds ratio for comparing $\Pr(Y = 1|P = p, X)$ and $\Pr(Y = 1|P = p + \theta, X)$ for specified values p and $\theta \neq 0$ is derived as

$$OR_{Y(P).X} = \frac{odds_{Y(P=p+\theta,X)}}{odds_{Y(P=p,X)}} = \frac{e^{(\beta_0+\beta_1\cdot(p+\theta)+\beta_2\cdot X)}}{e^{(\beta_0+\beta_1\cdot p+\beta_2\cdot X)}}$$

$$= e^{[(\beta_0-\beta_0)+(\beta_1\cdot p+\beta_1\cdot\theta-\beta_1\cdot p)+(\beta_2\cdot X-\beta_2\cdot X)]}$$

$$= e^{\beta_1\cdot\theta}$$

which represents an odds ratio related to a unit change by θ in P.

A logistic regression model employs Maximum Likelihood (ML) procedures to estimate parameters in the model. The ML is a general algorithm for estimating population parameters and the method can be applied to various situations. For instance, when an OLS regression is fitted to model normally distributed data, it yields regression coefficient estimators identical to the ML estimators fitted to the same data. One of the major advantages of the ML estimation is that it can estimate population parameters in nonlinear models, including logistic regression, which is of interest here. The goal of the method of ML is to estimate regression coefficients that maximize the probability of the observed data, or equivalently, maximize the "likelihood function."[89] We will briefly introduce

[89] Note that, in contrast, the OLS method seeks to minimize the sum of a squared difference between the observed

the principle of the ML estimation and hypothesis testing.

The probability of event Y for i^{th} individual given P_i and X_i is denoted by π_{P_i,X_i} for $i = 1, ..., N$. Let $Y_i = 0$ if the event Y occurred for i^{th} individual and $Y_i = 0$ otherwise. Then the likelihood function of the observed data is defined as

$$L(\beta) = \prod_i \pi_{P_i,X_i}^{Y_i} (1 - \pi_{P_i,X_i})^{1-Y_i}$$

where β represents the set of parameters, β_0, β_1, and β_2. Notice that the contribution of i^{th} individual to the likelihood function is π_{P_i,X_i} if $Y_i = 1$, and $1 - \pi_{P_i,X_i}$ if $Y_i = 0$. It is computationally easier to work with the log likelihood function — the natural log form of the likelihood function — which is expressed as

$$l(\beta) = \sum_i \{Y_i \cdot \ln[\pi_{P_i,X_i}] + (1 - Y_i) \cdot \ln[1 - \pi_{P_i,X_i}]\}$$

The task of estimating the parameters β_0, β_1, and β_2 is achieved by solving a system of three equations in three unknowns through the likelihood function. The task is achieved by differentiating the system of log likelihood functions with respect to β_j for $j = 0, 1, 2$ in our example. The ML estimates for β_0, β_1, and β_2 are obtained by setting each of the three equations equal to zero and finding the solution for each β_j as

$$\sum_i (Y_i - \pi_{P_i,X_i}) = 0$$

$$\sum_i P_i \cdot (Y_i - \pi_{P_i,X_i}) = 0$$

$$\sum_i X_i \cdot (Y_i - \pi_{P_i,X_i}) = 0$$

Our question in program evaluation is whether program participation P is related to the outcome Y, holding X constant in $\Pr(Y = 1|P, X) = \pi_{P,X} = \frac{1}{1+e^{[-(\beta_0+\beta_1 \cdot P+\beta_2 \cdot X)]}}$. The null and alternative hypotheses (H_0 and H_a) are defined as

$$H_0 : \beta_1 = 0|\beta_2$$

$$H_a : \beta_1 \neq 0|\beta_2$$

Alternatively, the hypotheses can be formulated in terms of an odds ratio. Recall that the odds ratio of the event Y for $P = 1$ relative to $P = 0$, holding X constant, is $OR_{Y(P).X} = e^{\beta_1}$ and $OR_{Y(P).X} = 1$ when $\beta_1 = 0$ (the value under the null hypothesis). Therefore, the hypotheses can be formulated as

$$H_0 : OR_{Y(P).X} = 1$$

$$H_a : OR_{Y(P).X} \neq 1$$

and predicted values of Y, that is, $\sum_i \left(Y_i - \hat{Y}_i\right)^2$.

There are multiple options to perform the hypothesis testing. We will provide a brief overview of the Wald test, which is one of the hypothesis testing methods about β_1 (or equivalently $OR_{Y(P).X}$) in the model.[90]

3.3.1.2.1 Wald Test

The Wald test is based on the Wald statistic, denoted by W, to compare the maximum likelihood estimate of the regression coefficient(s) of interest with its value under the null hypothesis. Specifically, under the null hypothesis in our example

$$H_0 : \beta_1 = 0|\beta_2$$

the probability of event Y conditional on P and X is modeled as

$$\Pr\left(Y = 1|P, X\right) = \frac{1}{1 + e^{[-(\beta_0+\beta_1\cdot P+\beta_2\cdot X)]}}$$

Then the Wald statistic is defined by

$$W = \frac{\hat{\beta}_1}{se\left(\hat{\beta}_1\right)}$$

where $se\left(\hat{\beta}_1\right) = \sqrt{V\left(\hat{\beta}_1\right)}$ obtained from the variance-covariance estimation in ML.[91] The Wald statistic W follows a standard normal distribution ($W \sim N\left(0,1\right)$) under the null.[92] In the Wald test, the null hypothesis is rejected at the significance level α when $|W| > c$ where c represents the critical value for the standard normal distribution (or equivalently, when $W^2 > c^2$ where c^2 is obtained from either the standard normal distribution or χ^2_1).

The Wald test can be extended to test multiple parameters simultaneously while holding a set of control variables constant. Let $\vec{P} = (P_1, ..., P_j)$ denote a set of regressors to test simultaneously

[90] Other major methods include the likelihood ration test and the Lagrange multiplier test (the score test). The likelihood ratio test is based on the difference in predicted values of Y between a model that includes the variable of interest and a model that excludes it. The likelihood ratio and Wald tests are both based on the likelihood function and asymptotically equivalent (Rao, 1965). Their test results in most cases are therefore very similar and we usually reach the same qualitative conclusions from these tests, while their performance can be affected by different conditions (for detailed discussion, see, for example, Peduzzi, Concato, Kemper, Holford, & Feinstein, 1996, and Hauck & Donner, 1977). The Lagrange multiplier test is yet another method that relies on the likelihood function as does the likelihood ratio test and approximates the likelihood ratio test. The Lagrange multiplier test however does not involve estimates of regression coefficients ($\hat{\beta}_j$) for computing the test statistic. We do not cover the likelihood ratio test or the Lagrange multiplier test in this section mainly because the theories of sample size estimation are well developed for the Wald test.

[91] The derivation of the variance-covariance estimation in ML is beyond the scope of the manual; however, interested readers are referred to Agresti (2013).

[92] Alternatively, W^2 can be expressed as

$$W^2 = \frac{\hat{\beta}_1^2}{V\left(\hat{\beta}_1\right)}$$

Then W^2 follows a chi-squared distribution with one degree of freedom ($W^2 \sim \chi^2_1$) under the null. This is derived from the relationship between the standard normal and χ^2 distributions in which $Z^2 = \chi^2_1$.

for $j \geq 2$ and $\vec{X} = (X_{j+1}, ..., X_k)$ for $k \geq 3$. Then the model is defined as

$$\Pr\left(Y = 1|\vec{P}, \vec{X}\right) = \frac{e^{[\beta_0+\beta_1\cdot P_1+\cdots\beta_j\cdot P_j+\beta_{j+i}\cdot X_{j+1}+\cdots+\beta_k\cdot X_k]}}{1+e^{[\beta_0+\beta_1\cdot P_1+\cdots\beta_j\cdot P_j+\beta_{j+i}\cdot X_{j+1}+\cdots+\beta_k\cdot X_k]}}$$

and the hypotheses are formulated as

$$H_0 : \beta_1 = \beta_2 = ... = \beta_j = 0|\beta_{j+i}, ..., \beta_k$$

$$H_a : \text{ at least one of the coefficients in } \beta_1, ..., \beta_j \text{ is not zero}|\beta_{j+i}, ..., \beta_k$$

Then the Wald statistic follows the χ^2 distribution with j degrees of freedom ($W \sim \chi^2_j$) under the null.[93] The null hypothesis is rejected at the significance level α when $W > c$ where c represents the critical value obtained from $\chi^2_{j,1-\alpha}$.

3.3.1.2.2 Sample Size Estimation for Logistic Regression

Hsieh (1989) and Hsieh, Bloch, and Larsen (1998) propose two methods of power and sample size estimations for a Wald-type test in logistic regression analysis: one for a logistic regression model with binary regressors, and the other for a logistic regression model with continuous regressors with a general exponential distribution. The method is applicable to a broad range of outcome probabilities and can account for a correlation between the regressor of interest and control variables in the model by introducing in the formula a "variance inflation factor." The variance inflation factor is a function of a multiple correlation coefficient between the regressor and control variables, that is, a "tolerance" of the regressor of interest for other regressors in the model.[94] This is particularly important for impact evaluation within the framework of selection on observables for parsing out program impact from any relationships between observed characteristics and the outcome. The method, however, is found to overestimate required sample size, especially in the case of multiple logistic regression analysis, and Hsieh (1989) and Hsieh, et al. (1998) propose it as a method for calculating an upper bound of required sample size, rather than the exact required sample size. Also the method for a continuous regressor can produce inaccurate results when the regressors have distributions other than normal or exponential distribution (or bivariate versions of these distributions in case of multiple logistic regression).

We will focus on the multiple logistic regression model of

$$logit\left[\Pr\left(Y = 1|P, X\right)\right] = \ln\left[odds\left(Y = 1|P, X\right)\right] = \ln\left[\frac{\pi_{(P,X)}}{1-\pi_{(P,X)}}\right] = \beta_0 + \beta_1 \cdot P + \beta_2 \cdot X$$

[93] The Wald statistic W for testing multiple parameters is computed as

$$W = \vec{\hat{\beta}}' \left[\vec{I}\left(\hat{\beta}\right)\right] \vec{\hat{\beta}}$$

where $\vec{\hat{\beta}}' = \left(\hat{\beta}_1, ..., \hat{\beta}_j\right)$ and $\vec{I}$ denotes the information matrix.

[94] Note that the same concept was introduced in multiple linear regression.

Then our null hypothesis is

$$H_0 : \beta_1 = 0|\beta_2$$

against the alternative hypothesis

$$H_a : \beta_1 \neq 0|\beta_2$$

As we discussed in the first section of Chapter 3, we will need to frame the alternative hypothesis in terms of a more specific threshold for violation $\beta_1 \neq 0|\beta_2$ that we wish to detect. This is accomplished by specifying the degree d_β to detect as

$$H_a : \beta_1 \geq d_\beta|\beta_2 \text{ or } \beta_1 \leq -d_\beta|\beta_2$$

Our natural question is: How do we specify d_β? We saw in multiple linear regression that the transformation between different parameters to specify the degree to detect required a substantial number of additional parameter specifications in some cases (especially in the case of the regression coefficient β_1). In multiple logistic regression, it is simpler and requires specifications of two parameters, namely $\Pr(Y = 1|P = p, X)$ and $\Pr(Y = 1|P = p + 1, X)$ for a specified value p. For instance, for a binary program variable P, our interest lies in comparing $\Pr(Y = 1|P = 0, X)$ and $\Pr(Y = 1|P = 1, X)$.

Recall that

$$e^{\beta_1} = OR_{Y(P).X} = \left[\frac{\Pr(Y = 1|P = p + 1, X)}{1 - \Pr(Y = 1|P = p + 1, X)}\right] / \left[\frac{\Pr(Y = 1|P = p, X)}{1 - \Pr(Y = 1|P = p, X)}\right]$$

which implies

$$\beta_1 = \ln\left[OR_{Y(P).X}\right] = \ln\left[\left(\frac{\Pr(Y = 1|P = p + 1, X)}{1 - \Pr(Y = 1|P = p + 1, X)}\right) / \left(\frac{\Pr(Y = 1|P = p, X)}{1 - \Pr(Y = 1|P = p, X)}\right)\right]$$

Then d_β can be determined with priors for $\Pr(Y = 1|P = p, X)$ and $\Pr(Y = 1|P = p + 1, X)$ as

$$d_\beta = \ln\left[\frac{\pi_{(p+1,X)}}{1 - \pi_{(p+1,X)}} / \frac{\pi_{(p,X)}}{1 - \pi_{(p,X)}}\right]$$

The sample size estimation method proposed by Hsieh (1989) and Hsieh, et al. (1998) takes two steps:

1. Estimate the sample size required to test for β_1 in simple logistic regression, i.e., in the absence of X. Call it n_s.
2. Inflate the estimated sample size by dividing n_s by $1 - R^2_{P(X)}$ (the tolerance of P for X) to obtain the estimated sample size required to test for β_1 in multiple logistic regression, that is, in the presence of X.

Note that the tolerance of P for X plays a role in multiple logistic regression as it does in multiple linear regression. Because $1 - R^2_{P(X)}$ is contained in $[0, 1]$, the sample size required for the hypothesis testing in multiple logistic regression is always larger than the sample size required in simple logistic regression when $R^2_{P(X)} \neq 0$. Moreover, the lower is the tolerance (the higher is $R^2_{P(X)}$), the larger is the increase in the required sample size.

3.3.1.2.3 Sample Size Estimation for a Binary Regressor P_b

First we will discuss the sample size estimation method proposed by Hsieh, et al. (1998) for hypothesis testing about a binary regressor P_b accounting for a regressor X. The first step of the sample size estimation for a binary regressor P_b is to estimate n_s, the sample size required to test for β_1 in a simple logistic regression model of

$$logit\left[\Pr\left(Y=1|P_b\right)\right] = \beta_0 + \beta_1 \cdot P_b$$

The sample size n_s as proposed by Hsieh, et al. (1998) is estimated based on the formula offered by Fleiss, et al. (2003) for a two-proportion comparison test, which we discussed in the first section of this chapter[95]

$$n_0 = \frac{\left[z_{1-\alpha/2} \cdot \sqrt{(r+1\cdot)\,\pi \cdot (1-\pi)} + z_{1-\beta} \cdot \sqrt{r \cdot \pi_{(0,X)} \cdot \left(1-\pi_{(0,X)}\right) + \pi_{(1,X)} \cdot \left(1-\pi_{(1,X)}\right)}\right]^2}{r \cdot \left(\pi_{(1,X)} - \pi_{(0,X)}\right)^2}$$

where $n_1 = n_0$ and $n_2 = r \cdot n_1$ are the sample size from populations with $P=0$ and $P=1$, respectively, for $r = \frac{n_2}{n_1}$; $\pi_{(0,X)}$ and $\pi_{(1,X)}$ denote $\Pr\left(Y=1|P=0,X\right)$, and $\Pr\left(Y=1|P=1,X\right)$, respectively, and π denotes $\Pr\left(Y=1\right)$. Then the sample size n_s in a simple logistic regression setting is

$$n_s = n_1 + n_2 = (1+r) \cdot n_0$$

$$= (1+r) \cdot \left\{ \frac{\left[z_{1-\alpha/2} \cdot \sqrt{(r+1\cdot)\,\pi \cdot (1-\pi)} + z_{1-\beta} \cdot \sqrt{r \cdot \pi_{(0,X)} \cdot \left(1-\pi_{(0,X)}\right) + \pi_{(1,X)} \cdot \left(1-\pi_{(1,X)}\right)}\right]^2}{r \cdot \left(\pi_{(1,X)} - \pi_{(0,X)}\right)^2} \right\}$$

The second step is to inflate the estimated sample size n_s to account for a regressor X in the multiple regression model[96]

$$logit\left[\Pr\left(Y=1|P_b\right)\right] = \beta_0 + \beta_1 \cdot P_b + \beta_2 \cdot X$$

[95] Note that we have changed the notation in this section to focus on the modeling of the probability of event in a multiple logistic regression setting. That is, π is analogous to $\overline{\overline{Y}}$, and $\pi_{(0,X)}$ and $\pi_{(1,X)}$ are analogous to $\overline{Y}_0$ and $\overline{Y}_1$ in the first section of Chapter 3.

[96] The modification of n_s can be extended to a model with multiple regressors to control, say $X_1, ..., X_q$ for $q > 1$. Let $\vec{X}$ denote a set of regressors as $\vec{X} = \{X_1, ..., X_q\}$. Then the tolerance of P for a set of regressors to control is defined as

$$1 - R_{P_b(\vec{X})} = 1 - \frac{SS_{REG[P_b(\vec{X})]}}{SS_{TOTAL[P_b]}}$$

This is achieved by dividing n_s by $1 - R^2_{P_b(X)}$ (the tolerance of P_b for X), where

$$1 - R^2_{P_b(X)} = 1 - \frac{SS_{REG[P_b(X)]}}{SS_{TOTAL[P_b]}}$$

The estimator of sample size required to test for β_1 accounting for X in multiple logistic regression is defined by

$$n = \frac{n_s}{1 - R^2_{P_b(X)}}$$

While a multiple correlation coefficient is typically applicable to continuous regressors, Hsieh, et al. (1998) demonstrated that this adjustment works adequately when the regressor of interest is a binary variable.

The method is best demonstrated through a numerical example. Suppose we wish to detect $\pi_{(1,X)} - \pi_{(0,X)} = 0.1$ for $\pi_{(0,X)} = 0.2$ and $\pi_{(1,X)} = 0.3$. Further suppose we assume $r = 1.5$ (that is, n_2 and n_1 are 60% and 40% of n, respectively) and a prior for π as $\pi = \frac{n_1}{n} \cdot \pi_{(0,X)} + \frac{n_2}{n} \cdot \pi_{(1,X)} = 0.4 \cdot 0.2 + 0.6 \cdot 0.3 = 0.26$. Then n_0 at $\alpha = 0.05$ (two-sided) and $1 - \beta = 0.8$ is 247 because

$$n_0 = \frac{\left[1.959963985\sqrt{(1.5+1)\cdot 0.26 \cdot (1-0.26)} + 0.841621234\sqrt{1.5 \cdot 0.2 \cdot (1-0.2) + 0.3 \cdot (1-0.3)}\right]^2}{1.5 \cdot (0.1)^2}$$

$$= 246.7576989 \approx 247$$

Then the sample size n_s is

$$n_s = (1 + r) \cdot n_0 = (1 + 1.5) \cdot 247 = 617.5 \approx 618$$

Now that we have estimated n_s, the next step is to inflate it by the level of tolerance. Suppose that $R^2_{P(X)}$ takes a value of 0.2, resulting in the tolerance of 0.8. Then the sample size required to test for β_1 accounting for X in multiple logistic regression is derived as

$$n = \frac{n_1}{1 - R^2_{P_b(X)}} = \frac{618}{1 - 0.2} = 772.5 \approx 773$$

We conclude that the estimated sample size is 773.

Note that the statistical power $1 - \beta$ can be estimated by working backwards from n to n_s, and from n_s to n_0 to solve for $z_{1-\beta}$ in the sample size formula for n_0. Given a sample size of n, n_s is derived as

$$n_s = n \cdot (1 - R^2_{P(X)})$$

Then n_0 is derived as

$$n_0 = \frac{n_s}{(1 + r)}$$

Finally, $z_{1-\beta}$ is derived by solving for $z_{1-\beta}$ from the formula for n_0 as

$$z_{1-\beta} = \frac{\sqrt{n_0 \cdot r} \cdot \left(\pi_{(1,X)} - \pi_{(0,X)}\right) - z_{1-\frac{\alpha}{2}} \cdot \sqrt{(r+1) \cdot \pi \cdot (1-\pi)}}{\sqrt{r \cdot \pi_{(0,X)} \cdot (1-\pi_{(0,X)}) + \pi_{(1,X)} \cdot \left(1 - \pi_{(1,X)}\right)}}$$

Then power is estimated as

$$1 - \beta = \Pr\left(Z > -z_{1-\beta}\right)$$

for a specified value $z_{1-\beta}$.

The power and sample size estimations proposed by Hsieh, et al. (1998) are based on a closed formula and straightforward for manual calculation. However, we can take advantage of the STATA command `power twoproportions` to estimate n_s, the sample size for a simple logistic regression, and facilitate the process.

The basic syntax for the command `power twoproportions` is:

```
 power twoproportions p1 p2 , power(numlist) alpha(numlist)
n(numlist) nratio(numlist)
```

where `p1` and `p2` represent $\pi_{(0,X)}$ and $\pi_{(1,X)}$, respectively. Then the program allows for specification of other parameters: the statistical significance level (α [two-sided]) in `alpha`; the statistical power level $(1-\beta)$ in `power`; the sample size (n) in `n`; and the ratio r in `nratio`. We specify a prior for each parameter in the parentheses. For instance, a significance level of 0.05 is specified as `alpha(0.05)` in the command line.[97]

Let us confirm our manual computation of the sample size that we performed for the parameter specifications of $\pi_{(0,X)} = 0.2$, $\pi_{(1,X)} = 0.3$, and $r = 1.5$ at $\alpha = 0.05$, and $1 - \beta = 0.8$. Then `power twoproportions` is specified:

```
 power twoproportions 0.2 0.3 , power(0.8) alpha(0.05) nratio(1.5)
```

which produces STATA Output 3.14. Notice that `n(numlist)` is left out of the command line so that STATA recognizes that the required output is the estimated sample size.

Below the first line "`Estimated sample sizes for a two-sample proportions test`" in the output, we see the null and alternative hypotheses in `Ho: p2 = p1 versus Ha: p2 != p1`. Then the parameter specifications are provided: α in `alpha`, $1-\beta$ in `power`, $\pi_{(1,X)} - \pi_{(0,X)}$ in `delta`, $\pi_{(0,X)}$ in `p1`, $\pi_{(1,X)}$ in `p2`, and r in `N2/N1`. The estimated sample sizes of those with $P = 0$ and $P = 1$ are denoted by `N1` and `N2`, respectively. The estimated total sample size is denoted by N, which is 618 and equals our manual computation of $n_s = 618$.[98] Then the final sample size n, which take into account X in multiple logistic regression, can be estimated manually by diving n_s by $1 - R^2_{P(X)}$, i.e., $n = \frac{n_s}{1-R^2_{P(X)}}$.

[97] The STATA program `power twoproportions` allows for both sample size and statistical power estimations. The program recognizes which output is required by reading which parameters are specified by a user and which parameter specification is left unspecified. When estimating power, a sample size should be specified in `n(numlist)` while power as in `power(numlist)` is left out of the command line. When estimating a sample size, a sample size `n(numlist)` should be left out of the command.

[98] The fractional sample size can be obtained by specifying `nfractional` in the command line.

STATA Output 3.14

```
Estimated sample sizes for a two-sample proportions test
Pearson´s chi-squared test
Ho: p2 = p1  versus  Ha: p2 != p1

Study parameters:

        alpha =    0.0500
        power =    0.8000
        delta =    0.1000  (difference)
           p1 =    0.2000
           p2 =    0.3000
        N2/N1 =    1.5000

Estimated sample sizes:

            N =       618
           N1 =       247
           N2 =       371
        N2/N1 =    1.5020
```

3.3.1.2.4 Sample Size Estimations for a Continuous Regressor P_c

Next we discuss the sample size estimation method proposed by Hsieh (1989) for hypothesis testing on a continuous regressor P_c accounting for a regressor X. The first step is to estimate n_s, the sample size required to test for β_1 in simple logistic regression of

$$logit\left[\Pr\left(Y=1|P_c\right)\right] = \beta_0 + \beta_1 \cdot P_c$$

Because of the nonlinear relationship between Y and P_c, the incremental change in $\Pr\left(Y=1|P_c\right)$ is not uniform for different levels of P_c. This implies that the difference to detect needs to be specified at a specific level of P_c.

The method proposed by Hsieh (1989) is designed to estimate the sample size required to test β_1 when the degree to detect is defined as a incremental change in $\Pr\left(Y=1|P_c\right)$ from $\overline{P}_c$ to $\overline{P}_c + S_{P_c}$ where σ_{P_c} denotes a standard deviation of P_c. Let $\pi_{\left(\overline{P},X\right)}$ and $\pi_{\left(\overline{P}+S,X\right)}$ denote $\Pr\left(Y=1|\overline{P}_c, X\right)$ and $\Pr\left(Y=1|\overline{P}_c+S_{P_c}, X\right)$, respectively. Notice that the odds ratio is then related to a unit change by σP_c in P_c as

$$OR_{Y(P_c).X} = \frac{odds_{Y\left(P_c=\overline{P}_c+S_{P_c},X\right)}}{odds_{Y\left(P_c=\overline{P}_c,X\right)}} = \frac{e^{\left(\beta_0+\beta_1\cdot(\overline{P}_c+S_{P_c})+\beta_2\cdot X\right)}}{e^{\left(\beta_0+\beta_1\cdot\overline{P}_c+\beta_2\cdot X\right)}} = e^{\beta_1\cdot S_{P_c}}$$

This implies that the degree to detect, d_β, needs to reflect a unit change by S_{P_c} in P_c as well for power and sample size estimations. Let $d_{S,\beta}$ denote the degree to detect for a unit change by S_{P_c}

in P_c as

$$d_{S,\beta} = \ln\left[\frac{\pi_{(\overline{P}+S,X)}}{1-\pi_{(\overline{P}+S,X)}} \Big/ \frac{\pi_{(\overline{P},X)}}{1-\pi_{(\overline{P},X)}}\right]$$

and its relationship with d_β, which corresponds to a one-unit change in P_c, is

$$d_{S,\beta} = S_{P_c} \cdot d_\beta$$

The sample size n_s as proposed by Hsieh (1989) is estimated as

$$n_s = \frac{\left[z_{1-\alpha} + z_{1-\beta} \cdot e^{\left(\frac{-d^2_{S,\beta}}{4}\right)}\right]^2 \cdot \left(1 + 2 \cdot \pi_{(\overline{P},X)} \cdot \delta\right)}{\pi^*_{(\overline{P},X)} \cdot d^2_{S,\beta}}$$

where

$$\delta = \frac{1 + \left[\left(1 + d^2_{S,\beta}\right) \cdot e^{\left(\frac{5 \cdot d^2_{S,\beta}}{4}\right)}\right]}{1 + e^{\left(\frac{-d^2_{S,\beta}}{4}\right)}}$$

Note that the formula estimates a sample size for one-sided hypothesis testing.

The second step is to inflate the estimated sample size n_s to account for a regressor X in the multiple regression model of[99]

$$logit\left[\Pr\left(Y=1|P_c\right)\right] = \beta_0 + \beta_1 \cdot P_c + \beta_2 \cdot X$$

Again, this is achieved by dividing n_s by $1 - R^2_{P(X)}$, i.e., the tolerance of P_c for X, where

$$1 - R^2_{P_c(X)} = 1 - \frac{SS_{REG[P_c(X)]}}{SS_{TOTAL[P_c]}}$$

The estimator of a sample size required to test for β_1 accounting for X in multiple logistic regression is defined by

$$n = \frac{n_s}{1 - R^2_{P(X)}}$$

Again we will demonstrate the steps through a numerical example. Suppose we want to perform hypothesis testing on β_1 for parameter specifications of $\pi_{(\overline{P},X)} = 0.2$ and $\pi_{(\overline{P}+S,X)} = 0.2$ at $\alpha = 0.05$ (one-sided) and $1 - \beta = 0.8$, which implies that

$$d_{S,\beta} = \ln\left[\frac{0.3}{1-0.3} \Big/ \frac{0.2}{1-0.2}\right] = \ln\left[1.714285714\right] = 0.538996501$$

[99] Again, the modification of n_s can be extended to a model with multiple regressors to control for $\vec{X} = \{X_1, ..., X_q\}$ by employing the squared multiple regression coefficient of P_c on $\vec{X}$.

and

$$\delta = \frac{1+\left[(1+(0.538996501)^2)\cdot e^{\left(\frac{5(0.538996501)^2}{4}\right)}\right]}{1+e^{\left(\frac{-(0.538996501)^2}{4}\right)}} = 1.479609524$$

Then n_s at the level of $\alpha = 0.05$ (one-sided) and $1-\beta = 0.8$ is estimated as

$$n_s = \frac{\left[1.644853627 + 0.841621234 \cdot e^{\left(\frac{-(1.479609524)^2}{4}\right)}\right]^2 \cdot (1+2\cdot 0.2\cdot 1.479609524)}{0.2\cdot(1.479609524)^2} = 161.4442524 \approx 162$$

The estimated sample size is 162. Now that we have estimated n_s, the next step is to inflate it by the level of tolerance. Suppose $R^2_{P_c(X)} = 0.2$. Then the sample size required to test for β_1 accounting for X in multiple logistic regression is derived as

$$n = \frac{n_1}{1-R^2_{P_c(X)}} = \frac{162}{1-0.2} = 202.5 \approx 203$$

Note that the statistical power $1-\beta$ for a continuous regressor P_c can be estimated by working backwards from n to n_s to solve for $z_{1-\beta}$ in the sample size formula for n_s. Given a sample size of n, n_s is derived as

$$n_s = n\cdot(1-R^2_{P_c(X)})$$

Then $z_{1-\beta}$ is derived as

$$z_{1-\beta} = \left[\sqrt{\frac{n_s\cdot\pi^*_{(\overline{P},X)}\cdot d^2_{S,\beta}}{1+2\cdot\pi^*_{(\overline{P},X)}\cdot\delta}} - z_{1-\alpha}\right] \Big/ \left[e^{\frac{-d^2_{S,\beta}}{4}}\right]$$

Finally the statistical power is estimated as

$$\Pr(Z > -z_{1-\beta})$$

for the derived value $z_{1-\beta}$.

In STATA, the sample size estimation for logistic regression as proposed by Hsieh (1989) can be performed using a user-written program `powerlog`.[100]

The basic syntax of the command `powerlog` is:

```
powerlog , p1(value) p2(value) [ rsq(value) alpha(value) ]
```

where `p1` and `p2` represent $\pi_{(\overline{P},X)}$ and $\pi_{(\overline{P}+S,X)}$, respectively. Then the program allows for specifications of other parameters: the statistical significance level (α [one-sided]) in `alpha`, and $R^2_{P(X)}$ in `rsq`. We specify a prior for each parameter in the parentheses.

[100] The program was written by Philip B. Ender at Statistical Computing and Consulting, UCLA Academic Technology Services, and can be downloaded and installed by using the findit command, i.e., `findit powerlog`. The Institute for Digital Research and Education, UCLA, provides on its website examples of how to use the `powerlog` program.

Let us confirm our manual computation of the sample size that we performed for the parameter specifications of $\pi_{(\overline{P},X)} = 0.2$, $\pi_{(\overline{P}+S,X)} = 0.3$, and $R^2_{P_c(X)} = 0.2$, at $\alpha = 0.05$ (one-sided), and $1 - \beta = 0.8$. The command `powerlog` is specified:

```
powerlog, p1(0.2) p2(0.3) alpha(.05) rsq(0.2)
```

which produces STATA Output 3.15.

STATA Output 3.15

```
Logistic regression power analysis
One-tailed test: alpha=.05  p1=.2  p2=.3  rsq=.2  odds ratio=1.714285714285714

power          n
0.60         121
0.65         137
0.70         156
0.75         177
0.80         202
0.85         233
0.90         276
```

Note that `powerlog` estimates the sample size for a set of statistical powers ranging from 0.60 to 0.90 in an increment of 0.05. For a statistical power of 0.80, the estimated sample size is 202, which equals our estimation of 203 with a difference of 1 unit due to rounding.[101]

3.3.1.2.5 Logistic Regression Model vs. Linear Probability Model

One of the technical challenges related to logistic regression is in interpreting the results in terms of the estimated probabilities, which are often of primary interest. The regression coefficient β_1 in logistic regression represents the natural log of the odds ratio, and the incremental change in the probability of event Y associated with a one-unit change in P is not uniform across different levels of the base probability $\Pr(Y = 1|P = p)$. Let us examine this claim through a numerical example of a simple logistic regression with a binary regressor P, which is modeled as

$$logit\,[\Pr(Y = 1|P_b)] = \beta_0 + \beta_1 \cdot P$$

Recall that the relationship of Y with P in the simple logistic regression framework can be defined as

$$\Pr(Y = 1|P) = \frac{1}{1 + e^{[-(\beta_0 + \beta_1 \cdot P)]}}$$

[101] The command `powerlog` does not allow for a direct estimation of statistical power for a specified sample size; however, one can get an idea about the level of statistical power by examining the estimated sample size for each level of power. For instance, an estimated power for a sample size of 185 (which is between $n = 177$ for a power of 0.75 and $n = 202$ for a power of 0.80) falls between 0.75 and 0.80.

which implies

$$\Pr(Y=1|P=0)=\frac{1}{1+e^{[-(\beta_0)]}}$$

and

$$\Pr(Y=1|P=1)=\frac{1}{1+e^{[-(\beta_0+\beta_1)]}}$$

Suppose $\Pr(Y=1|P=0)=0.1$ which implies $\beta_0=-\ln(9)=-2.197224577$. Also suppose $\beta_1=1$. What would be $\Pr(Y=1|P=1)$ and the incremental change in the probability of event Y associated with a change in P from $P=0$ to $P=1$? We solve the equation of $\Pr(Y=1|P=1)$ as

$$\Pr(Y=1|P=1)=\frac{1}{1+e^{[-(\beta_0+\beta_1)]}}=\frac{1}{1+e^{[-(-2.197224577+1)]}}\approx 0.2320$$

Then the incremental change in the probability of event Y is

$$\Pr(Y=1|P=1)-\Pr(Y=1|P=0)=0.2320-0.1=0.1320$$

Now let us suppose $\Pr(Y=1|P=0)=0.2$ which implies $\beta_0=-\ln(4)=-1.386294361$, while $\beta_1=1$ remains the same. What would be $\Pr(Y=1|P=1)$ and the incremental change in the probability of event Y associated with a change in P from $P=0$ to $P=1$? Again, we solve the equation of $\Pr(Y=1|P=1)$ as

$$\Pr(Y=1|P=1)=\frac{1}{1+e^{[-(\beta_0+\beta_1)]}}=\frac{1}{1+e^{[-(-1.386294361+1)]}}\approx 0.4046$$

Then the incremental change in the probability of event Y is

$$\Pr(Y=1|P=1)-\Pr(Y=1|P=0)=0.4046-0.2=0.2046$$

which does not equal 0.1320, the incremental change from $\Pr(Y=1|P=0)=0.1$ with the same level of β_1 at 1.

Because of the nonlinear relationship between the incremental change in the probability of event Y associated with a change in P and the base probability $\Pr(Y=1|P=p)$, it is challenging to interpret the regression coefficient β_1 in terms of the magnitude of the incremental change in the probability of event Y, which is often of primary interest in impact evaluation. One approach to addressing the technical challenge is to evaluate marginal effects, which represent the predicted change in the probability of event Y associated with a change in a regressor of interest. Marginal effects can be evaluated at specified values of regressors, for instance, average, median, specific values of interest, etc. However, note that the predicted marginal effects depend on a specified level of regressors — that is, the incremental change in the probability of event Y is still conditional on the base probability $\Pr(Y=1|P=p)$.

Another approach is to apply the multiple linear regression method to model the relationship between a binary outcome Y and regressors P and X as

$$Y = \beta_0 + \beta_1 \cdot P + \beta_2 \cdot X + \varepsilon$$

When a binary outcome Y is modeled within the framework of linear regression, it is called a "linear probability model" and is designed to describe the relationship between the probability of event Y and regressors as

$$\Pr(Y = 1|P, X) = \beta_0 + \beta_1 \cdot P + \beta_2 \cdot X$$

That is, the regression coefficient β_1 for P represents an incremental change in the probability of event Y associated with a one-unit change in P while accounting for X. The approach using a linear probability model has both pros and cons. One major disadvantage is that the predicted probability is not guaranteed to be contained in [0,1], the range of a probability. This is particularly so when the observed $\overline{Y}$ of a sample is close to 0 or 1. Another disadvantage is that the variance of Y conditional on P and X varies across values of P and X. This implies that a linear probability model violates the homoscedasticity assumption in OLS, on which t- and F-tests that we described are based. Wooldridge (2006) recommends estimation of heteroscedasticity-robust standard errors through Weighted Least Squares (WLS) estimation to address the problem.[102] On the other hand, the major advantage of employing a linear probability model over a logistic regression model is that the interpretation of the estimated regression coefficient for P is straightforward: it is the estimated incremental change in the probability of event Y associated with a one-unit change in P. Moreover, the estimated incremental change is modeled to be uniform across different levels of the base probability $\Pr(Y = 1|P = p)$.

3.3.2 Matching

Matching is another method within the framework of "selection on observables." Matching estimates program impact for each individual by finding a similar individual who experienced the counterfactual outcome. For a participant the counterfactual outcome is Y^0, while for a non-participant it is Y^1. For each individual, their counterfactual outcome is estimated by the outcome experienced by a similar person for whom that counterfactual is observed. Thus, the counterfactual for a non-participant would be the outcome $Y (= Y^0)$ for a similar participant. In this manner, an estimate of $Y^1 - Y^0$ can be formed for each observed individual. It is an estimate because the value of either Y^1 or Y^0 will have been estimated for each individual according to their participation status. With the estimates of program impact $Y^1 - Y^0$ so formed for each observed individual, forming an estimate of average impact for whatever population the observed individuals represent simply involves suitably averaging across them. Within the general method of matching, there are different approaches to finding a similar individual who experienced the counterfactual outcome, or to forming an estimate of average impact. In this section, we will review two major methods

[102] The technical details are beyond the scope of the manual. Interested readers are encouraged to read "7.5 A Binary Dependent Variable: the Linear Probability Model" and "8.5 The Linear Probability Model Revisited" in Wooldridge (2006).

of matching: exact matching and propensity score matching. Furthermore, within the propensity score matching method, we will review two approaches: nearest neighbor matching and propensity score weighting. However, the research and literature on matching are evolving and readers should be aware that what we are providing here is by no means a comprehensive coverage of methods or approaches to matching. Interested readers are referred to Stuart (2010).

3.3.2.1 Exact Matching

Exact matching seeks to estimate the program impact from a sample of participants and non-participants matched on each of their observed characteristics that a researcher wishes to balance between the two groups. That is, in exact matching, participants and non-participants are matched on every combination of distinct values of the observed covariates of interest. For instance, suppose one wishes to match participants and non-participants on sex (male and female) and region of residence (urban and rural), that is, two covariates with two distinct values each. This example yields $2 \times 2 = 4$ combinations of the covariate values as shown in the following two-by-two contingency table in Table 3.6.

Table 3.6. Two by two contingency table

	Sex: Male	**Sex: Female**
Region: Urban	$Male - Urban$	$Female - Urban$
Region: Rural	$Male - Rural$	$Female - Rural$

The procedure is to classify participants and non-participants into one of the combinations of the covariate values that are mutually exclusive and collectively exhaustive. In this example, observations are classified into four combinations of sex and residential type: male-urban, female-urban, male-rural, and female-rural. Then the average outcome within each combination is calculated, followed by estimation of the program impact as a weighted average of the combination-specific outcome differences. The most frequently employed weights are those based on the frequency of observations.

Although the procedure of exact matching is rather simple, exact matching may be impractical when there are a large number of covariates, or if they are continuous rather than discrete (while this may be addressed through variants or combinations of matching methods, for instance, the Mahalanobis method with propensity score and coarsened exact matching), resulting in a high dimension of covariate vector, that is, a large number of combinations of covariate values. This problem is known as the "curse of dimensionality." For instance, suppose we decide to add three more covariates to the previous example with four combinations of sex and region of residence: age with 15 categories defined by five-year intervals (for example, 15–19, 20–24, and so forth), educational attainment with 5 categories, and income with 5 categories. Then the covariate dimension suddenly grows from $2 \times 2 = 4$ to $2 \times 2 \times 15 \times 5 \times 5 = 1,500$ combinations of the covariate categories. Aside from its implacability due to the large number of combinations required on which to match participants and non-participants, a high dimension of covariate combinations may result in insufficient numbers of observations within the combinations. This increases the likelihood of not being able to form counterfactual matches because everyone in a particular category either always participated or never participated. When individuals without a counterfactual match are dropped from the analysis of program impact evaluation, there are two immediate statistical implications. First, it results in a smaller sample size available for the analysis and hence a lower statistical power than expected with the original sample size. Second, because certain types of individuals are excluded from the analysis, it can result in a biased estimate of the program impact (Rosenbaum & Rubin, 1985).

3.3.2.2 Propensity Score Matching

The propensity score method was developed by Rosenbaum and Rubin (1983) to address the limited application of exact matching arising from the "curse of dimensionality." The approach for reducing the dimensionality is to match participants to a non-participant who experienced the counterfactual outcome and had a similar probability of participation conditional on observed characteristics. This probability of participation conditional on a vector of observed characteristics is called a propensity score and is estimated for each individual. That is, the propensity score matching avoids the problem of high dimensionality by mapping the covariate vector into the one-dimensional measure through scalar functions such as the probit and logit functions. Various methods and approaches have been proposed within the framework of propensity score matching since the introduction of the concept. In this section we will briefly review two approaches: k:1 nearest neighbor matching and propensity score weighting.

3.3.2.2.1 K:1 Nearest Neighbor Matching

Once a propensity score is estimated for each participant and non-participant based on one of the aforementioned procedures, differences in these propensity scores among any participants and non-participants can be measured. K:1 nearest neighbor matching seeks to create matches of participants and non-participants of similar propensity scores by matching participants with non-participants who have the smallest difference in their propensity scores from the participants. The number of non-participants to be assigned to each participant can be controlled by the researcher. When only one non-participant is selected for each participant, the method is called 1:1 nearest neighbor matching in which, for each participant, the non-participant with the smallest difference in the propensity score from the participant is selected. When k non-participants (for $k > 1$) are selected for each participant, the top k non-participants with the smallest difference in the propensity score from the participant are selected. In either method of nearest neighbor matching, non-participants that were not selected as matches based on their difference in the propensity scores from participants are discarded from the sample, resulting in the aforementioned statistical implications: a smaller sample size available for the analysis (and hence a lower statistical power than expected with the original sample size), and a potentially biased estimate of the program impact. While selecting multiple non-participants per participant may seem attractive at first for controlling the variance due to the larger analytical sample size (the number of observations available for analysis) compared to selecting only one non-participant per participant, it should be noted that there could a trade-off between bias and variance. In theory, when multiple non-participants per participant are selected, their overall difference in the propensity scores from the participant is expected to be greater compared to when the closest non-participant is selected. This larger overall difference in the propensity scores can lead to an increased bias in the program impact estimate. While the technical discussion of different methods is beyond the scope of this manual, interested readers are referred to Stuart (2010).

3.3.2.2.2 Propensity Score Weighting

Another way of applying the estimated propensity scores is to use them as weights in the analysis for estimating the program impact. There are mainly three weighting methods: inverse probability of program participation, odds of program participation, and kernel weighting. While the technical details of the different methods are beyond the scope of this manual, the overall purpose of propensity score weighting is to weigh participants and non-participants through a function of the estimated propensity scores in order to reduce bias in the program impact estimate.[103]

3.3.2.2.3 Power and Sample Size Estimations for Matching

To our knowledge, there is no estimator or formula developed explicitly to estimate power or sample size for performing matching analysis. This is not surprising given that there are multiple and different tools under the broad approach of matching. For instance, in our brief and limited review, we have studied three different methods: exact matching, propensity score matching, and propensity score weighting. Given that they employ different technical procedures, it is not feasible to have a single estimator that can be applied to different tools.

However, research on matching has been evolving, and there have been studies on power or sample size estimation for matching methods in recent years. Jung, Chow, and Chi (2007) proposed a power and sample size estimation method employing a weighted Mantel-Haenszel test statistic proposed by Gart (1985). A Mantel-Haenszel test was originally developed by Mantel and Haenszel (1959) to examine an association between a binary outcome and a binary independent variable. It is closely related to a χ^2 test for one contingency table, but a Mantel-Haenszel allows for hypothesis testing that adjusts for a potentially confounding and observed variable. In a Mantel-Haenszel test the confounding variable is treated as a stratifying variable. Within each stratum defined by the confounding variable, a contingency table of the outcome and independent variables is created, allowing for different associations between the outcome and independent variables across strata. Then a Mantel-Haenszel test derives an overall odds ratio from the combination of contingency tables and performs hypothesis testing based on a test statistic that is approximately a chi-square random variable with one degree of freedom. Nam (1998) applied the Mantel-Haenszel test to develop a power and sample size calculation method for propensity score matching, treating propensity

[103] The weighting methods yield different interpretations of the estimated program impact. Weighting by the inverse probability of program participation seeks to yield a sample representative of both participants and non-participants, while weighting by the odds yields a sample representative of participants.

scores as the confounding variable in the original Mantel-Haenszel test. That is, units are grouped into strata defined by ranges of estimated propensity scores so that within each strata units have similar propensity scores. Jung, et al. (2007) extended the work by Nam (1998) and generalized weights, which are a function of the allocation proportion of units by the status of the outcome and independent variables within each stratum, to employ a weighted Mantel-Haenszel test statistic proposed by Gart (1985).

We discuss briefly the method proposed by Jung, et al. (2007) for applying a weighted Mantel-Haenszel test to matching. Suppose we are interested in a binary outcome Y, for $Y = 1$ if the event is observed, and $Y = 0$ otherwise. Suppose we have two groups of interest, denoted by P, where $P = 1$ indicates Group 1, and $P = 2$ indicates Group 2. Suppose there are H strata defined by ranges of estimated propensity scores. Then the data can be organized within a framework of H number of 2×2 contingency tables. The h^{th} contingency table is presented in Table 3.7. The number of individuals of the i^{th} row and j^{th} column in h^{th} stratum is denoted by n_{ijh} (for $i = 1, 2$, $j = 1, 2$, and $h = 1, ..., H$), which defines $n_{\cdot jh} = n_{1jh} + n_{2jh}$, $n_{i \cdot h} = n_{i1h} + n_{i2h}$, $n_{\cdot\cdot h} = n_{11h} + n_{12h} + n_{21h} + n_{22h}$, and $n = \sum_{h=1}^{H} n_{\cdot\cdot h}$.

Table 3.7. h^{th} **Two by two contingency table**

	P =1	**P=2**	**Total**
Y=1	n_{11h}	n_{12h}	$n_{1\cdot h}$
Y=0	n_{21h}	n_{22h}	$n_{2\cdot h}$
Total	$n_{\cdot 1h}$	$n_{\cdot 2h}$	$n_{\cdot\cdot h}$

Let p_{jh} denote the probability of event Y among j^{th} group in h^{th} stratum, $\Pr(Y = 1|P = j)$. Then the null and alternative hypotheses are formulated as

$$H_0 : p_{1h} = p_{2h}$$

and

$$H_a : p_{1h} \neq p_{2h}$$

For the significance level of α (two-sided) and the statistical power $1 - \beta$, the sample size estimator proposed by Jung, et al. (2007) is given as

$$n = \frac{\left(\sigma_0 \cdot z_{1-\frac{\alpha}{2}} + \sigma \cdot z_{1-\beta}\right)^2}{\delta^2}$$

where

$$\sigma_1 = \sqrt{\sum_{h=1}^{H} w_h^2 \cdot a_h \cdot b_{1h} \cdot b_{2h} \cdot [b_{2h} \cdot p_{1h} \cdot (1 - p_{1h}) + b_{1h} \cdot p_{2h} \cdot (1 - p_{2h})]}$$

$$\sigma_0 = \sqrt{\sum_{h=1}^{H} w_h^2 \cdot a_h \cdot b_{1h} \cdot b_{2h} \cdot (b_{1h} \cdot p_{1h} + b_{2h} \cdot p_{2h}) \cdot [b_{1h} \cdot (1 - p_{1h}) + b_{2h} \cdot (1 - p_{2h})]}$$

for

$$a_h = \frac{n_{\cdot\cdot h}}{n}$$

$$b_{jh} = \frac{n_{\cdot jh}}{n_h}$$

$$p_{jh} = \Pr\left(Y = 1 | P = j\right)$$

$$\delta = \sum_{h=1}^{H} w_h \cdot a_h \cdot b_{1h} \cdot b_{2h} \cdot \left(p_{1h} - p_{2h}\right)$$

and w_h, the weight employed in the weighted Mantel-Haenszel test statistic, is estimated as

$$\hat{w}_h = \frac{n_{2\cdot h}}{n_h}$$

Then the statistical power is estimated as

$$1 - \beta = \Pr\left[Z > \left(\frac{\sigma_1}{\sigma_0}\right) \cdot z_{1-\frac{\alpha}{2}} - \sqrt{n} \cdot \left(\frac{|\delta|}{\sigma_1}\right)\right]$$

The proposed method provides a closed-form solution to the power and sample calculation and is straightforward in theory; however, it has major limitations when performed in practice. The sample size calculation based on the weighted Mantel-Haenszel statistic requires inputs for a relatively large number of parameters, and many, if not all, of them may not have prior information or data on which to base assumptions. For instance, researchers may find it challenging to determine the allocation proportion of the participant and comparison groups within each stratum, which may depend on the distribution of the propensity scores. Similarly, the expected outcome needs to be specified for participant and comparison groups within each stratum. Given that the sample size calculation involves assumptions on a relatively large number of parameters, it is recommended that a sensitivity analysis of the calculated sample size be conducted based on various combinations of the parameter inputs. Additionally, the weighted Mantel-Haenszel test relies on the asymptotic distribution of the test statistic and its performance under small sample size needs to be assessed. Also note that the estimators are designed to estimate power and sample size for a binary outcome variable.

In the absence of robust estimators designed to estimate power or sample size specifically for matching analysis, estimators designed for other study designs have often been employed in practice. One commonly-adopted approach to estimating power and sample size for matching analysis is to employ a method designed for a blocked randomized design. This approach is based on the assumption of matching analysis: selection on observables. That is, participant and comparison groups differ only in observables, and once they are controlled for, the program impact can be estimated. With this assumption, the matching variables or propensity scores play a similar role as a block in a blocked randomized design. Then the power and sample size estimation methods discussed in the section on randomized controlled trials in Chapter 3 are employed for matching analysis. Another approach is to employ power and sample size estimation methods designed for multiple linear regression analysis as discussed in the section on regression in Chapter 3, treating the propensity score as the regressor of interest.

However, one must be cautioned that there are technical challenges unique to the matching approach in estimating power or sample size. First, when the distribution of the estimated propensity scores differs between the participant and comparison groups, it results in a limited region of common support and requires a researcher to discard units outside the region of common support at the analysis stage. This ultimately reduces the sample size available for analysis and hence undermines the statistical power. The same concept can be extended to exact matching in that there are cells with only program or comparison units. One strategy to address the problem is to inflate the sample size by the expected proportion of units in the region of common support to account for the loss of analysis units. However, the distribution of the covariates are typically study- or content-specific, and unknown prior to data collection, making it difficult to make an assumption about the loss of analysis units due to a limited region of common support.

Second, the propensity scores should be estimated from all relevant covariates that potentially confound the program impact, which implies that the study should be designed to collect data on each of these covariates. When information is missing on any one or more of these covariates, the propensity score for the observation with a missing value cannot be estimated and therefore will be dropped from the analysis, reducing the sample size available for the analysis as well as the statistical power. While the problem can be regarded as a form of non-response, which is relevant and a concern to any study, the problem is potentially more serious for a propensity score analysis, as it requires non-missing data for all covariates for each individual. Furthermore, when the excluded individuals due to missing values are systematically different from other individuals in characteristics relevant to both the outcome and treatment status, the sample used to estimate the program is no longer representative of the entire sample and the estimated program impact may be biased.

3.4 Within Models

We briefly delve into the arena of "within" estimators of program impact.[104] Within estimators seek to identify program impact by examining whether *changes* in the outcome of interest appear to be caused by *changes* in program exposure. These changes occur within some overall unit of observation. For instance, we might examine how changes in an individual's program participation over time appear to cause changes in his/her outcomes of interest. They are called within estimators because they seek to identify program impact from variation in program participation and outcomes *within* some unit of observation, such as the individual. For the most part, we can apply what we have already discussed by way of sample size estimation strategies to arrive at (hopefully) reasonable sample size estimates for many within models.

However, we concede from the outset that there are some within models for which we can offer

[104] Much of the discussion in this section draws liberally from material in Lance, et al. (2014), particularly to set the stage for the discussions of sample size estimation.

no truly useful sample size estimation guidance. For instance, we know of no sample size estimation approach truly adequate to the Chamberlain model, developed for use with logit regression (see Maddala, 1983 for more details). In general, it can be seen as a kind of template for circumstances where we can offer little useful guidance: it is a relatively complicated within estimator developed for a fairly specific estimation setting (nonlinear limited dependent variable regression) that is complex in and of itself. Even in cases of relatively simple applications of within estimation to the nonlinear setting, one runs into relatively complicated issues for which there is no general guidance, either from sample size estimation theory or mainstream statistics or econometrics.

For instance, one alternative to the Chamberlain model, with its complex derivation and reliance on relatively unapproachable concepts for many practitioners (such as the whole idea of a "sufficient statistic") is the sometimes-employed, relatively straightforward strategy in nonlinear regression contexts such as the logit of simply including a dummy variable for the unit of observation (for example, Angeles, et al., 1998). For instance, confronted with a panel data set at the individual level, one might perform logit regression of the outcome of interest at each point of observation on program participation at the same point in time and a dummy variable for each individual in the data set. Unfortunately, this is subject to the "incidental parameters" problem (see Neyman & Scott; 1948, Green, 2004; Hsiao, 1986.), for which we can offer no general guidance. For instance, in the present hypothetical circumstance, we know of no general, formulaic way of establishing how many panel observations for each individual would be enough to insure that sampling goals for detecting impact would be met.

Nonetheless, the vast majority of applications of within models tend to involve relatively simple linear models, and for these some guidance is available. We begin our discussion with basic classic linear within estimators. For the present purposes we consider basically two categories of estimators: the first differences estimator and the "demeaned" estimator (the latter being what most would refer to as the "fixed effect" estimator).

First, some review. Let us start with a model of potential outcomes and the costs of participation. Although within estimators can be operationalized in a number of senses,[105] the usual conceptualization of the within variation is with respect to *time*. We will focus on variation over time for individuals from a sample of individuals observed at more than one point in time. We therefore need to introduce a time dimension to our variables. Thus, for instance, the outcome of interest is Y_{it}, signifying that this is the outcome of interest observed for individual i at time t.

We consider the following potential outcome equations:

$$Y_{it}^{0} = \beta_0 \cdot t + \beta_2 \cdot X_{1it} + \beta_3 \cdot X_{2i} + \beta_4 \cdot \mu_i + \epsilon_{it}^{Y}$$

$$Y_{it}^{1} = \beta_0 \cdot t + \beta_1 + \beta_2 \cdot X_{1it} + \beta_3 \cdot X_{2i} + \beta_4 \cdot \mu_i + \epsilon_{it}^{Y}$$

where X_{1it} is a time-varying observed individual characteristic, X_{2i} is a fixed observed individual characteristic, μ_i is a fixed unobserved individual characteristic, and ϵ_{it}^{Y} is time-varying unobserv-

[105] For instance, one might consider variation within a household, community, etc.

able.

The cost of participation is given by

$$C_{it} = \gamma_0 + \gamma_1 \cdot X_{1it} + \gamma_2 \cdot X_{2i} + \gamma_3 \cdot \mu_i + \epsilon^C_{it}$$

At any given time period t individual i choose to participate (i.e., P_{it} will equal 1) if

$$Y^1_{it} - Y^0_{it} - C_{it} > 0$$

or

$$\beta_1 - \gamma_0 - \gamma_1 \cdot X_{1it} - \gamma_2 \cdot X_{2i} - \gamma_3 \cdot \mu_i - \epsilon^C_{it} > 0$$

There are a few important features to note about this. First, the observed individual characteristics X_{1it} and X_{2i} influence participation and hence we would expect that, other things being equal, the average values of these characteristics should differ between participants and non-participants. There will thus clearly be selection into program participation based on observables. Second, the same is true of the fixed unobservable μ_i. It influences the participation decision and hence would be expected to differ in terms of average value between participants and non-participants. Third, by dint of the fact that they are fixed, characteristics such as X_2 and μ cannot explain variation in program participation (or the outcome) *over time* for the i^{th} individual.

Let us now derive a regression specification. The observed outcome is

$$\begin{aligned} Y_{it} &= P_{it} \cdot Y^1_{it} + (1 - P_{it}) \cdot Y^0_{it} \\ &= P_{it} \cdot \left(\beta_0 \cdot t + \beta_1 + \beta_2 \cdot X_{1it} + \beta_3 \cdot X_{2i} + \beta_4 \cdot \mu_i + \epsilon^Y_{it}\right) \\ &+ (1 - P_{it}) \cdot \left(\beta_0 \cdot t + \beta_2 \cdot X_{1it} + \beta_3 \cdot X_{2i} + \beta_4 \cdot \mu_i + \epsilon^Y_{it}\right) \\ &= \beta_0 \cdot t + \beta_1 \cdot P_{it} + \beta_2 \cdot X_{1it} + \beta_3 \cdot X_{2i} + \beta_4 \cdot \mu_i + \epsilon^Y_{it} \end{aligned}$$

It is already clear that unbiased estimation of program impact (i.e., unbiased estimation of β_1) might be a bit tricky. Unbiased estimation would require the independence or mean independence of program participation P_{it} and μ_i and the independence or mean independence of P_{it} and ϵ^Y_{it}. Per the latter condition, since ϵ^Y_{it} played no role in determining program participation, the only readily plausible way that the independence or mean independence of P_{it} and ϵ^Y_{it} could be violated is if there was some sort of relationship between ϵ^Y_{it} and ϵ^C_{it}. We rule out this possibility by assumption.

The sticky wicket is then presented by μ_i. It clearly plays a role in shaping the program participation decision and hence should be correlated with P_{it}. Thus, to recover an unbiased estimate of β_1, μ_i must either be controlled for or somehow purged from the regression specification that we actually estimate.

Within estimators purge μ_i from the regression equation that we will actually estimate. Suppose that we observed a sample of n individuals at two points in time, $t = 1$ and $t = 2$. Specifically, suppose that we observed $\{Y_{it}, P_{it}, X_{1it}, X_{2i}\}$ for $i = 1, ..., n$ individuals at times $t = 1, 2$.

Suppose (for simplicity) that there are only two time periods, $t = 1$ and $t = 2$. Subtracting the regression terms at $t = 1$ from those at $t = 2$ (in other words, taking the first difference between the regression terms in the two time periods), we have

$$\begin{array}{rrcccccccccccccc} & Y_{i2} & = & \gamma & + & \beta_0 \cdot 2 & + & \beta_1 \cdot P_{i2} & + & \beta_2 \cdot X_{1i2} & + & \beta_3 \cdot X_{2i} & + & \beta_4 \cdot \mu_i & + & \epsilon_{i2}^Y \\ - & Y_{i1} & = & \gamma & + & \beta_0 \cdot 1 & + & \beta_1 \cdot P_{i1} & + & \beta_2 \cdot X_{1i1} & + & \beta_3 \cdot X_{2i} & + & \beta_4 \cdot \mu_i & + & \epsilon_{i1}^Y \\ \hline & \Delta Y_i & = & 0 & + & \beta_0 & + & \beta_1 \cdot \Delta P_i & + & \beta_2 \cdot \Delta X_{1i} & + & 0 & + & 0 & + & \Delta\epsilon_i^Y \end{array}$$

yielding a final specification to be estimated of

$$\Delta Y_i = \beta_0 + \beta_1 \cdot \Delta P_i + \beta_2 \cdot \Delta X_{1i} + \Delta\epsilon_i^Y$$

where $\Delta Y_i = Y_{i2} - Y_{i1}$, $\Delta P_i = P_{i2} - P_{i1}$, $\Delta X_{1i} = X_{1i2} - X_{1i1}$, and $\Delta\epsilon_i^Y = \epsilon_{i2}^Y - \epsilon_{i1}^Y$.

Notice that the problematic term μ_i has been purged from this new, first-differenced specification. Notice as well, however, that it is pretty costly information-wise. Suppose, for instance, that we have a sample with n individuals, each of whom is observed T times, involving $n \cdot T$ individual observations. The first difference model involves the loss of n observations to implement the model. For instance, in our simple example where $T = 2$ (and hence involves $2 \cdot n$ individual observations), we are actually left to estimate the first difference model with just n observations. This loss of information, equal to one panel's worth of observations, is one of the drawbacks of this model.

Another within estimator, often referred to as "fixed effects," relies on a slightly different process called demeaning. Put simply, to demean the key covariates (the outcome Y, program participation P and any time-varying covariates X) each observation for each of the key covariates has the mean of that variable for the unit of observation subtracted from it.

To fix ideas, let us return to the problematic (from the standpoint of obtaining an unbiased estimate of program impact β_1) basic cross-sectional specification

$$Y_{it} = \beta_0 + \beta_1 \cdot P_{it} + \beta_2 \cdot X_{1it} + \beta_3 \cdot X_{2i} + \beta_4 \cdot \mu_i + \epsilon_{it}^Y$$

Define $\overline{Y}_i$ as follows:

$$\overline{Y}_i = \sum_{t=1}^{T} \frac{Y_{it}}{T}$$

where T is the number of time periods over which individual i is observed. In other words, $\overline{Y}_i$ is the mean value of the outcome Y for individual i across the T time periods over which that individual is observed. For simplicity, assumed that all $i = 1, ..., n$ individuals in our hypothetical sample are observed for each of time period captured in the sample. In other words, assume that we have balanced panels in our sample.

$\overline{Y}_i$ is simply the average of Y for the i^{th} individual in the sample. For instance, in the setting of two time periods considered in the last subsection, $t = 1, 2$, $\overline{Y}_i$ is

$$\overline{Y}_i = \frac{Y_{i1} + Y_{i2}}{2}$$

We can similarly imagine averages for all of the right hand side variables in the specification: $\overline{P}_i$, $\overline{X}_{1i}$, $\overline{X}_{2i}$, $\overline{\mu}_i$, and $\overline{\epsilon}_i^Y$. Because X_{2i} and μ_i do not vary for each individual, note that

$$\overline{X}_{2i} = X_{2i}$$

and

$$\overline{\mu}_i = \mu_i$$

The remaining variables do vary over time for the individual, and hence will not equal their averages.

Let us now consider the basic regression specification, but applied to the demeaned variables:[106]

$$Y_{it} - \overline{Y}_i = \beta_0 \cdot (1-1) + \beta_1 \cdot \left(P_{it} - \overline{P}_i\right) + \beta_2 \cdot \left(X_{1it} - \overline{X}_{1i}\right) + \beta_3 \cdot \left(X_{2i} - \overline{X}_{2i}\right) + \beta_4 \cdot (\mu_i - \overline{\mu}_i) + \epsilon_{it}^Y - \overline{\epsilon}_i^Y$$

However, since

$$\overline{X}_{2i} = X_{2i}$$

and

$$\overline{\mu}_i = \mu_i$$

we know that

$$\mu_i - \overline{\mu}_i = 0$$

and

$$X_{2i} - \overline{X}_{2i} = 0$$

Therefore, the demeaned regression specification reduces to

$$Y_{it} - \overline{Y}_i = \beta_0 \cdot (1-1) + \beta_1 \cdot \left(P_{it} - \overline{P}_i\right) + \beta_2 \cdot \left(X_{1it} - \overline{X}_{1i}\right) + \epsilon_{it}^Y - \overline{\epsilon}_i^Y$$

Finally, the term $\beta_0 \cdot (1-1)$ is motivated by the idea of demeaning the constant term, which is based on a column of 1s (and hence has a mean of 1, across individuals or even the entire sample). This leads to our final specification

$$Y_{it} - \overline{Y}_i = \beta_1 \cdot \left(P_{it} - \overline{P}_i\right) + \beta_2 \cdot \left(X_{1it} - \overline{X}_{1i}\right) + \epsilon_{it}^Y - \overline{\epsilon}_i^Y$$

[106] This new specification can easily be motivated in a manner analogous to first differencing. First, beginning with

$$Y_{it} = \beta_0 \cdot 1 + \beta_1 \cdot P_{it} + \beta_2 \cdot X_{1it} + \beta_3 \cdot X_{2i} + \beta_4 \cdot \mu_i + \epsilon_{it}^Y$$

we have

$$\overline{Y}_i = \beta_0 \cdot 1 + \beta_1 \cdot \overline{P}_i + \beta_2 \cdot \overline{X}_{1i} + \beta_3 \cdot \overline{X}_{2i} + \beta_4 \cdot \overline{\mu}_i + \overline{\epsilon}_i^Y$$

Subtracting the second equation from the first yields the specification in the main body of the text

$$\begin{array}{cccccccccccc}
 & Y_{it} & = & \beta_0 \cdot 1 & + & \beta_1 \cdot P_{it} & + & \beta_2 \cdot X_{1it} & + & \beta_3 \cdot X_{2i} & + & ... \\
- & \overline{Y}_i & = & \beta_0 \cdot 1 & + & \beta_1 \cdot \overline{P}_i & + & \beta_2 \cdot \overline{X}_{1i} & + & \beta_3 \cdot \overline{X}_{2i} & + & ... \\
\hline
 & Y_{it} - \overline{Y}_i & = & \beta_0 \cdot (1-1) & + & \beta_1 \cdot \left(P_{it} - \overline{P}_i\right) & + & \beta_2 \cdot \left(X_{1it} - \overline{X}_{1i}\right) & + & \beta_3 \cdot \left(X_{2i} - \overline{X}_{2i}\right) & + & ...
\end{array}$$

This specification has purged the confounding unobservable μ, as well as the non-time varying observable X_2.

Most of the implications of the first difference model carry over to this model as well. For instance, the model still involves the loss of the information from one time period's worth of observations, though the reason can seem a bit less straightforward in this case. To illustrate, let us focus on the regressor

$$Y_{it} - \overline{Y}_i$$

By definition

$$\sum_{t=1}^{T} \left(Y_{it} - \overline{Y}_i \right) = 0$$

In other words, for individual i the deviations of his/her value of Y from the average over time for Y sum to zero. This means, however, that one observation for individual i does not really provide truly independent information since it must insure that this condition holds. Focusing on the T^{th} observation (i.e., assuming that $\{Y_{i1}, Y_{i2}, ..., Y_{iT-1}\}$ are independently determined and hence provide $T-1$ truly independent units worth of information about variation in Y), it must be the case that

$$Y_{iT} = \overline{Y}_i - \sum_{t=1}^{T-1} \left(Y_{it} - \overline{Y_i} \right)$$

Therefore, Y_{iT} does not, from this perspective, provide truly independent information about variation in Y: its value is locked down by the need to insure that the condition

$$\sum_{t=1}^{T} \left(Y_{it} - \overline{Y}_i \right) = 0$$

holds. However, such a loss of one truly independent observation will occur for each individual, resulting in the loss of n units of information (i.e., one time period worth of information) from demeaning.

Estimation of this demeaned model is relatively simply. Define

$$\dot{Y}_{it} = Y_{it} - \overline{Y}_i$$

$$\dot{P}_{it} = P_{it} - \overline{P}_i$$

and

$$\dot{X}_{1it} = X_{1it} - \overline{X}_{1i}$$

These can be calculated by straightforward means: simply subtract from the value of the variable for individual i at time t that variable's mean over time for individual i. Estimation is then simply a question of regressing $\dot{Y}$ on $\dot{P}$ and $\dot{X}_1$.

Another variant on the fixed effects model is often referred to as the "dummy variable" fixed effects model. It is simple enough in conceptualization. It attempts to control for μ_i from

$$Y_{it} = \gamma + \beta_0 \cdot t + \beta_1 \cdot P_{it} + \beta_2 \cdot X_{1it} + \beta_3 \cdot X_{2i} + \beta_4 \cdot \mu_i + \epsilon_{it}^Y$$

by estimating it for each of the $i = 1, ..., n$ individuals in the sample. This is accomplished by regressing Y_{it} on P_{it}, t, X_{1it}, and a dummy variable for each individual. The regression model actually estimated is thus on the lines of

$$Y_{it} = \delta + \beta_0 \cdot t + \beta_1 \cdot P_{it} + \beta_2 \cdot X_{1it} + \sum_{j=2}^{N} \phi_j \cdot d_j + \epsilon_{it}^Y$$

where d_j is a dummy that equals 1 if $i = j$ (for instance, $d_{550} = 1$ if $i = 550$). The ϕs are parameters to be estimated in the regression. They estimate the "fixed effect" of being each individual in the sample. A dummy variable is not offered for the first individual in the sample $i = 1$ because a fixed effect for that individual could not be separately identified from a constant term. Rather, the constant term δ captures that individual's fixed effect, and the estimates of the fixed effect (i.e., the ϕs) for every other individual in the sample can be interpreted as the difference between their own fixed effect and the first individual's fixed effect. This is the reason that the constant term is now labelled δ and not γ — in the dummy variable model we are not estimating the original constant γ.

For all of these (the first difference, demeaned fixed effects and dummy variable fixed effects) the ultimate estimation model can be viewed as a linear regression model. In principle, this takes this into the linear regression sample size estimation framework of the last section. What is important to remember, however, is the loss of information associated with these models, particularly the first difference and demeaned fixed effects regression. To fix ideas, let us cast the sample size estimation just in terms of the number of individuals in the sample (i.e., there will be only one observation per individual for the first difference or demeaned fixed effects regression).[107] If we determine that we will need to have observations for n individuals via the sample size estimators of the last section, then we will have to have $n \cdot 2$ observations in our eventual sample (i.e., a sample size of n for the regression implies a sample size of $n \cdot 2$ where each of the n individuals are interviewed twice). In general, however many times individuals will appear in the regression (given the overall sample size estimate and the number of individuals one plans to interview), they must be observed on more time than that given the information loss associated with these models.

Another major consideration in any application where a panel design is considered is attrition. This can sometimes be considerable and must be factored into the determination of the overall sample size. For instance, suppose that we need to have two panel observations (i.e., each individual will be interviewed up to two times) and need to have a sample of 1,000 individual observations for, for instance, a demeaned fixed effects regression. Then we need in principle to have 2,000

[107] How exactly an overall sample size of n should be parsed between individuals interviewed per panel (n^p) and the number of panels p (such that $n^p \cdot p = n$) is, provided the number of panels $p > 1$, going to depend on parameters such as survey cost associated with alternative parsing schemes.

observations in our sample (i.e., 1,000 individuals interviewed twice each). Suppose, however, that we learn that we will likely encounter 50 percent attrition. Then we would actually need to select 2,000 individuals for interview in the first panel with the expectation that 1,000 will be reinterviewed for the second panel. So to have 1,000 individual observations in our regression, we would need 3,000 individual-time observations and 2,000 longitudinal observations (from the 1,000 individuals who were successfully interviewed twice).

Let us compare OLS and within estimators from the point of view of sample size estimation. For the within estimation, we will focus on an example in which $T = 2$ (i.e., each individual is observed twice) with a first difference model

$$\Delta Y_i = \beta_0 + \beta_1 \cdot \Delta P_i + \beta_2 \cdot \Delta X_i + \Delta\epsilon_i^Y$$

where $\Delta Y_i = Y_{i2} - Y_{i1}$, $\Delta P_i = P_{i2} - P_{i1}$, $\Delta X_i = X_{i2} - X_{i1}$, and $\Delta\epsilon_i^Y = \epsilon_{i2}^Y - \epsilon_{i1}^Y$.

The null and alternative hypotheses are

$$H_0 : \beta_1 = 0|\beta_2$$

and

$$H_0 : \beta_1 \neq 0|\beta_2$$

A likely OLS analysis is focused on the data observed at $t = 2$, which reflects a situation in which only the endline data (or data collected after the program initiation) are available. The OLS model of our example is specified

$$Y_{2i} = \alpha_0 + \alpha_1 \cdot P_{2i} + \alpha_2 \cdot X_{2i} + \epsilon_{2i}$$

The null and alternative hypotheses are

$$H_0 : \alpha_1 = 0|\alpha_2$$

and

$$H_0 : \alpha_1 \neq 0|\alpha_2$$

In the previous section on the sample size estimator for OLS, we demonstrated the effect of parameter specifications, including σ_Y, σ_P, and $R^2_{P(X)}$, on the estimated power and sample size. For the OLS model focused on the endline data, the parameters of interest are σ_{Y_2}, σ_{P_2}, and $R^2_{P_2(X_2)}$. In the first difference model, on the other hand, the parameters need to be specified in terms of the difference in variables between the two time points, namely, $\sigma_{\Delta Y}$, $\sigma_{\Delta P}$, and $R^2_{\Delta P(\Delta X)}$. We will discuss the implication of the parameter specifications.

Note that

$$\begin{aligned}\sigma^2_{\Delta Y} &= V\left(Y_2 - Y_1\right) = \sigma^2_{Y_1} + \sigma^2_{Y_2} - 2 \cdot cov\left(Y_1, Y_2\right) \\ &= \sigma^2_{Y_2} + \left(\sigma^2_{Y_1} - 2 \cdot \rho_Y \cdot \sigma_{Y_1} \cdot \sigma_{Y_2}\right)\end{aligned}$$

where ρ_Y denotes the correlation coefficient between Y_1 and Y_2. This implies that

$$\sigma^2_{\Delta Y} \neq \sigma^2_{Y_2}$$

unless

$$\sigma^2_{Y_1} - 2 \cdot \rho_Y \cdot \sigma_{Y_1} \cdot \sigma_{Y_2} = 0$$

Specifically:

1. $\sigma^2_{\Delta Y} < \sigma^2_{Y_2}$ when $\sigma^2_{Y_1} - 2 \cdot \rho_Y \cdot \sigma_{Y_1} \cdot \sigma_{Y_2} < 0$
2. $\sigma^2_{\Delta Y} = \sigma^2_{Y_2}$ when $\sigma^2_{Y_1} - 2 \cdot \rho_Y \cdot \sigma_{Y_1} \cdot \sigma_{Y_2} = 0$
3. $\sigma^2_{\Delta Y} > \sigma^2_{Y_2}$ when $\sigma^2_{Y_1} - 2 \cdot \rho_Y \cdot \sigma_{Y_1} \cdot \sigma_{Y_2} > 0$

Because $\sigma_{Y_1} \geq 0$ and $\sigma_{Y_2} \geq 0$, this suggests that, for fixed σ_{Y_1} and σ_{Y_2}, ρ_Y drives the difference between σ_{Y_2} and $\sigma_{\Delta Y}$. While ρ_Y most likely takes a positive value, a larger ρ_Y increases the covariance between Y_1 and Y_2 and reduces $\sigma^2_{\Delta Y}$. This can be demonstrated in a numerical example. Suppose, for instance, $\sigma_{Y_1} = 3$, $\sigma_{Y_2} = 4$, and $\rho_Y = 0.3$ to compare $\sigma_{\Delta Y}$ and σ_{Y_2}. Then

$$\sigma^2_{\Delta Y} = \sigma^2_{Y_2} + \left(\sigma^2_{Y_1} - 2 \cdot \rho_Y \cdot \sigma_{Y_1} \cdot \sigma_{Y_2}\right)$$

$$= 4^2 + \left(3^2 - 2 \cdot 0.3 \cdot 3 \cdot 4\right) = 17.8 > 16 = \sigma^2_{Y_2}$$

Suppose now that $\rho_Y = 0.6$ to assess a change in $\sigma_{\Delta Y}$.

$$\sigma^2_{\Delta Y} = 4^2 + \left(3^2 - 2 \cdot 0.6 \cdot 3 \cdot 4\right) = 10.6 < 16 = \sigma^2_{Y_2}$$

The same argument can be applied to the other variables, namely P and X. Higher ρ_P and ρ_X, which denote a correlation between P_1 and P_2 and a correlation between X_1 and X_2, respectively, reduce $\sigma_{\Delta P}$ and $\sigma_{\Delta X}$, respectively, where

$$\sigma^2_{\Delta P} = V\left(P_2 - P_1\right) = \sigma^2_{P_1} + \sigma^2_{P_2} - 2 \cdot cov\left(P_1, P_2\right)$$

$$= \sigma^2_{P_2} + \left(\sigma^2_{P_1} - 2 \cdot \rho_P \cdot \sigma_{P_1} \cdot \sigma_{P_2}\right)$$

and

$$\sigma^2_{\Delta X} = V\left(X_2 - X_1\right) = \sigma^2_{X_1} + \sigma^2_{X_2} - 2 \cdot cov\left(X_1, X_2\right)$$

$$= \sigma^2_{X_2} + \left(\sigma^2_{X_1} - 2 \cdot \rho_X \cdot \sigma_{X_1} \cdot \sigma_{X_2}\right)$$

The specification of $R^2_{P(X)}$, which determines the tolerance level of P for X, also differs between the OLS and within estimator models. In the OLS model focused on the endline data it is derived by

$$R^2_{P_2(X_2)} = \left[\frac{cov\left(P_2, X_2\right)}{\sigma_{P_2} \cdot \sigma_{X_2}}\right]^2$$

whereas in the first difference model it is derived by

$$R^2_{\Delta P(\Delta X)} = \left[\frac{cov\left(\Delta P, \Delta X\right)}{\sigma_{\Delta P} \cdot \sigma_{\Delta X}}\right]^2$$

We already know the relationships between σ_{P_2} and $\sigma_{\Delta P}$, and between σ_{X_2} and $\sigma_{\Delta X}$, which leads us to focus on the covariance terms. $cov\left(P_2, X_2\right)$ in the OLS models is

$$cov\left(P_2, X_2\right) = \frac{\sum_{i=1}^{N}\left(P_{2i} - \overline{P}_2\right)\left(X_{2i} - \overline{X}_2\right)}{N}$$

whereas $cov\left(\Delta P, \Delta X\right)$ in the first difference model is derived by

$$\begin{aligned} cov\left(\Delta P, \Delta X\right) &= \frac{\sum_{i=1}^{N}\left(\Delta P_i - \overline{\Delta P}\right)\left(\Delta X_i - \overline{\Delta X}\right)}{N} \\ &= \frac{\sum_{i=1}^{N}\left[\left(P_{i2} - \overline{P}_2\right) - \left(P_{i1} - \overline{P}_1\right)\right]\left[\left(X_{i2} - \overline{X}_2\right) - \left(X_{i1} - \overline{X}_1\right)\right]}{N} \\ &= cov\left(P_2, X_2\right) + cov\left(P_1, X_1\right) - cov\left(P_2, X_1\right) - cov\left(P_1, X_2\right) \end{aligned}$$

where $\overline{\Delta P} = \frac{\sum_{i=1}^{N} \Delta P_i}{N}$ and $\overline{\Delta X} = \frac{\sum_{i=1}^{N} \Delta X_i}{N}$. This implies that

$$cov\left(P_2, X_2\right) \neq cov\left(\Delta P, \Delta X\right)$$

<u>unless</u>

$$cov\left(P_1, X_1\right) - cov\left(P_2, X_1\right) - cov\left(P_1, X_2\right) = 0$$

It is therefore crucial to understand the differences in the parameter specifications between OLS and within estimator models when estimating power or sample size. Again, note that the within estimator model with $T = 2$ requires 2 observations per unit.[108]

[108] The parameter specifications for a demeaned model can be derived in the same manner. Let $\dot{Y}_i = Y_{i2} - \overline{Y}_i$, $\dot{P}_i = P_{i2} - \overline{P}_i$, and $\dot{X}_i = X_{i2} - \overline{X}_i$. Then

$$\sigma^2_{\dot{Y}} = \frac{1}{4}\left(\sigma^2_{Y_2} + \sigma^2_{Y_1} - 2 \cdot \rho_Y \cdot \sigma_{Y_1} \cdot \sigma_{Y_2}\right)$$

$$\sigma^2_{\dot{P}} = \frac{1}{4}\left(\sigma^2_{P_2} + \sigma^2_{P_1} - 2 \cdot \rho_P \cdot \sigma_{P_1} \cdot \sigma_{P_2}\right)$$

$$\sigma^2_{\dot{X}} = \frac{1}{4}\left(\sigma^2_{X_2} + \sigma^2_{X_1} - 2 \cdot \rho_X \cdot \sigma_{X_1} \cdot \sigma_{X_2}\right)$$

$$\begin{aligned} cov\left(\dot{P}, \dot{X}\right) &= \frac{\sum_{i=1}^{N}\left(\dot{P}_i - \overline{\dot{P}}\right)\left(\dot{X}_i - \overline{\dot{X}}\right)}{N} \\ &= \frac{\sum_{i=1}^{N} \frac{1}{2}\left[\left(P_{i2} - \overline{P}_2\right) - \left(P_{i1} - \overline{P}_1\right)\right] \cdot \frac{1}{2}\left[\left(X_{i2} - \overline{X}_2\right) - \left(X_{i1} - \overline{X}_1\right)\right]}{N} \\ &= \frac{1}{4}\left[cov\left(P_2, X_2\right) + cov\left(P_1, X_1\right) - cov\left(P_2, X_1\right) - cov\left(P_1, X_2\right)\right] \end{aligned}$$

where

$$\overline{\dot{P}} = \frac{\sum_{i=1}^{N} \dot{P}_i}{N}$$

Finally, we discuss the sample size implications of one particular within model: the difference-in-differences (DID for brevity) model. This extremely popular impact model can be a deliberate design from the outset and is also a popular fallback position when an alternative panel design does not work out (such as failed randomization in an RCT for which baseline, pre-program observations were collected for samples of participants).

The basic "bare bones" difference-in-differences estimator is very simple. At a minimum, one must observe participants and non-participants at two points in time, before a program is introduced and after it is introduced (we will refer to these two points as time periods $t = 1$ and $t = 2$, respectively). Although we observe participants and non-participants both before and after the initiation of a program, it is critical that they can be empirically differentiated even *before* the program commences. Define

$$\overline{Y}_t^P$$

as the average (across some suitable representative sample) of an outcome of interest for participants at time period t (where $t = 1, 2$ in our simple example). Define

$$\overline{Y}_t^N$$

analogously for non-participants. The difference-in-differences estimate of program impact is then

$$\left(\overline{Y}_2^P - \overline{Y}_1^P\right) - \left(\overline{Y}_2^N - \overline{Y}_1^N\right)$$

From this expression, the origin of the term "difference-in-differences" should be evident: the estimator is the difference in the differences over time (between the pre- and post-program implementation phases) for the participant and non-participant groups.

The difference-in-differences estimator of program impact is thus the difference in changes over time in outcomes between participants and non-participants. In other words, any difference in changes over time between participants and non-participants can be ascribed to the program. But this implies that the changes over time experienced by non-participants are indicative of the changes participants would have experienced had they not begun participating after time period 1. The assumption that the changes in the outcome over time among non-participants indicate the changes that would have occurred among participants had they not participated after time period 1 (or, put more simply, that had they not participated, participants would have experienced the same average change in the outcome over time that non-participants did experience) is called the "parallel trend assumption."

The basic logic of this estimator is illustrated in Figures 3.18 – 3.21. We begin with Figure 3.18, which illustrates the trajectory of an outcome Y (specifically the trajectory of the average value

and

$$\overline{\dot{X}} = \frac{\sum_{i=1}^{N} \dot{X}_i}{N}$$

Again, this demonstrates that the parameter specifications differ between OLS and within estimator models when estimating power or sample size.

of Y) between two points in time, $t = 1$ and $t = 2$, for participants and non-participants. Between $t = 1$ and $t = 2$ the program is introduced. Hence when we distinguish between participants and non-participants, we do so with reference to the participation decision made *after* the introduction of the program; no one can participate *before* the program is introduced. Notice that the non-participant sample experienced a lower average outcome over time than participants, but there was some upward trend in the average outcome even for non-participants.

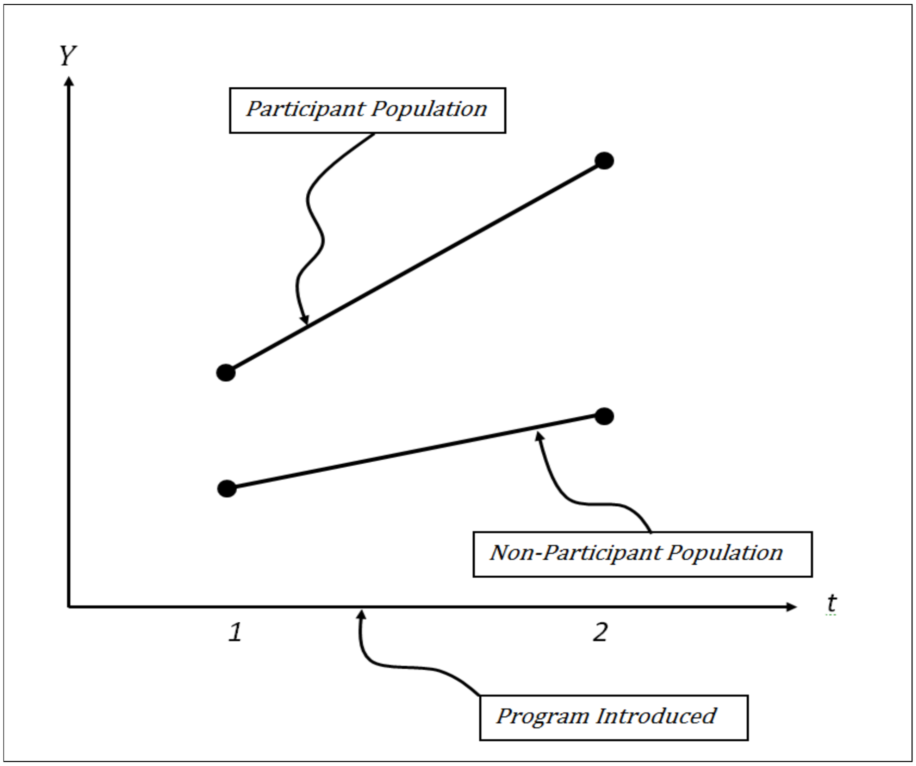


Figure 3.18. Outcome trajectories, participants and non-participants

Figure 3.19 makes more explicit the average outcomes for participants and non-participants over time. Following the discussion above, the average outcome is denoted

$$\overline{Y}_t^k$$

where $t = 1, 2$ are the two time periods and $k = P, N$ denote participants and non-participants, respectively. A simple and straightforward estimator of program impact might be

$$\overline{Y}_2^P - \overline{Y}_2^N$$

This estimator is simply the comparison of average outcomes between participants and non-participants at time $t = 2$. We can see from this figure that this might be a problematic estimator of program impact. Given that participants and non-participants exhibited different average outcomes at time period $t = 1$ (with $\overline{Y}_1^P > \overline{Y}_1^N$), it is not at all clear that the average level of Y

non-participants experienced at time period $t = 2$ is indicative of what participants would have experienced at $t = 2$ in the absence of the program. Moreover, there is no reason to believe that the experiences of participants are indicative of what non-participants would have experienced had they participated. There is thus no reason to believe that this is an unbiased estimator of the program impact.

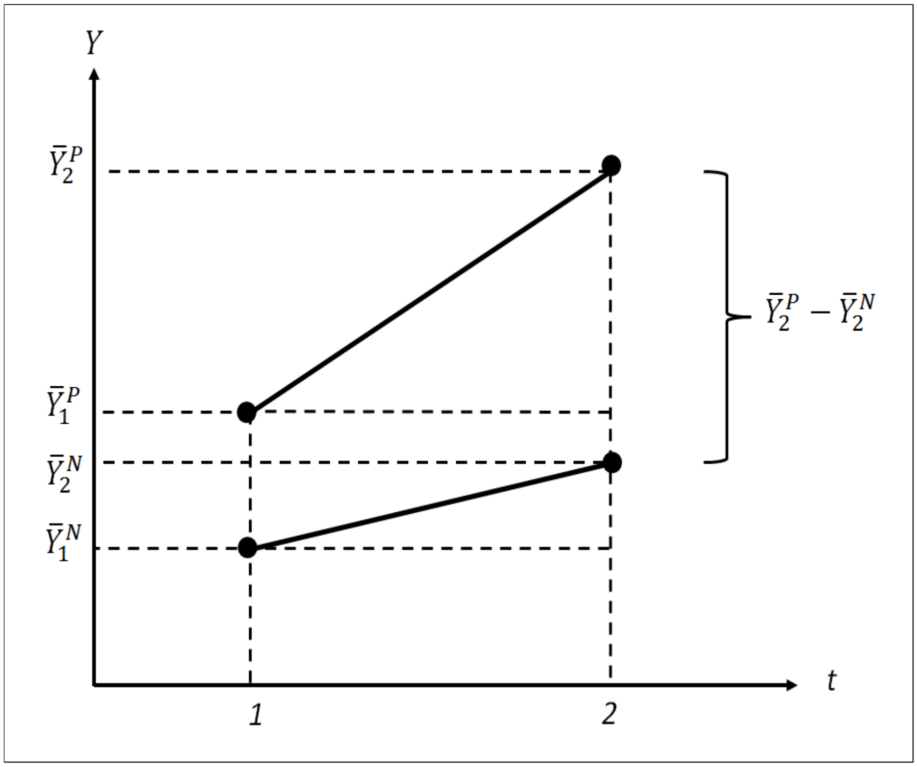


Figure 3.19. Changes over time, participants and non-participants

Figure 3.20 illustrates the parallel trend assumption. As we have seen, the solid lines between the dots represent the actual evolution of the average outcome for participants and non-participants. The double dotted line illustrates the parallel trend assumption. This is the assumed evolution of the outcome for participants had they not participated. Notice that it is parallel to the line indicating the actual change in the average outcome for non-participants. Thus, it assumes that the *change over time* for participants would, had they not begun participating after time period 1, have been the same as that for non-participants. The parallel nature of the lines is also the origin of the expression “parallel trend.”

The change over time represented by the parallel trend assumption is represented in Figure 3.20 by Δ. The parallel trend assumption is that

$$\Delta = \overline{Y}_2^N - \overline{Y}_1^N$$

In other words, the changes implied by the parallel trend assumption equal those that non-participants actually did experience.

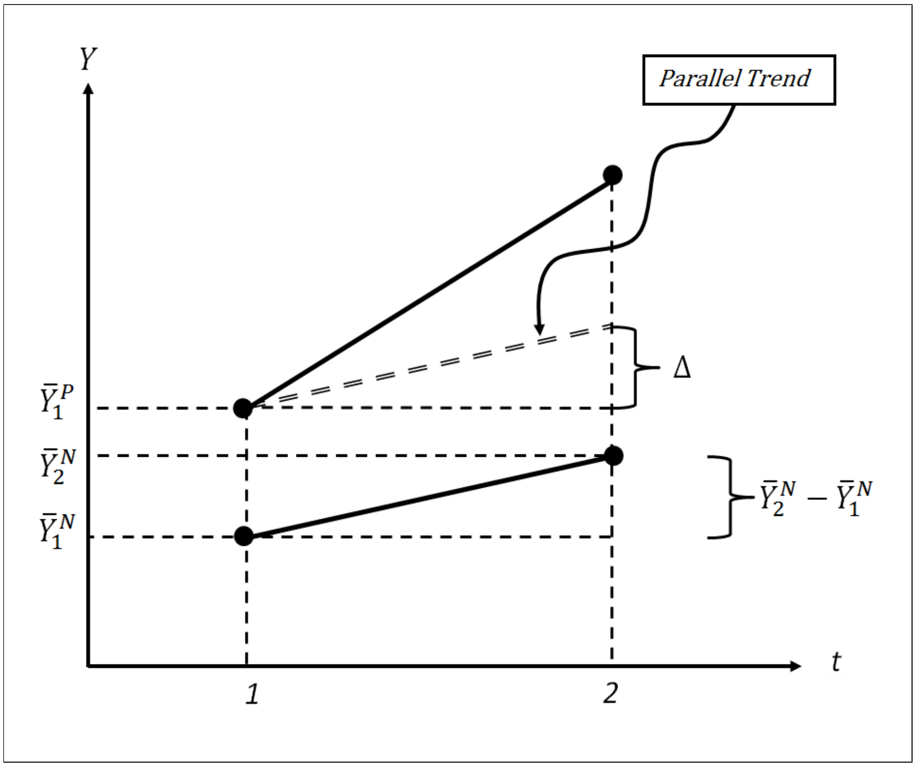


Figure 3.20. The parallel trend assumption

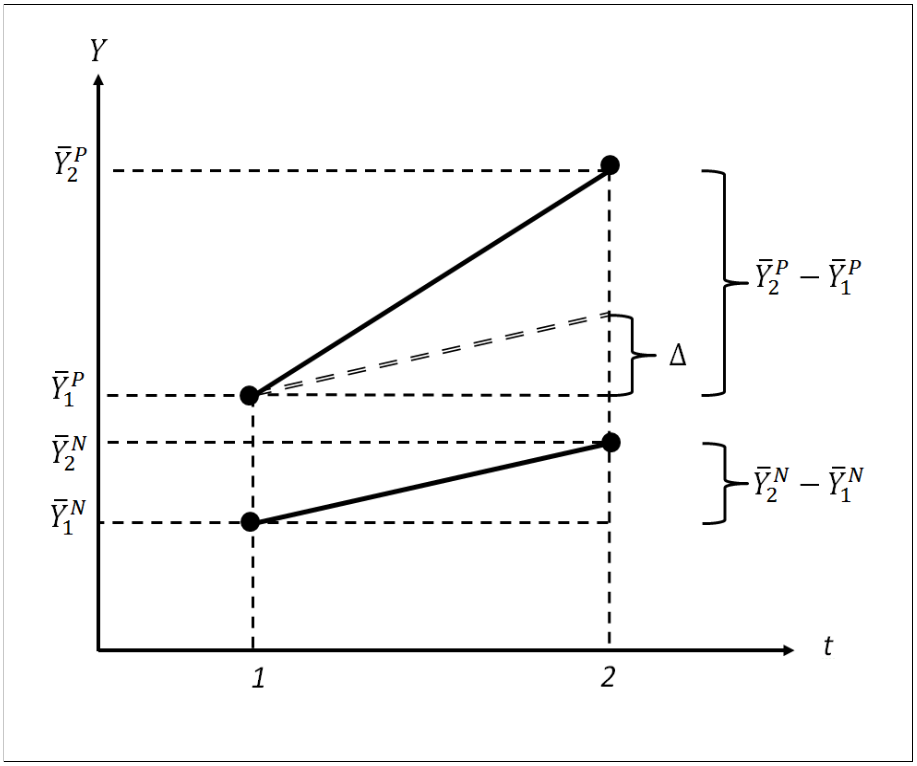


Figure 3.21. The difference-in-differences estimation

Finally, Figure 3.21 rounds out the graphical discussion. The difference-in-differences estimator is

$$\left(\overline{Y}_2^P - \overline{Y}_1^P\right) - \left(\overline{Y}_2^N - \overline{Y}_1^N\right)$$

However, the parallel trend assumption tells us that

$$\Delta = \overline{Y}_2^N - \overline{Y}_1^N$$

Hence

$$\left(\overline{Y}_2^P - \overline{Y}_1^P\right) - \left(\overline{Y}_2^N - \overline{Y}_1^N\right)$$

$$\overline{Y}_2^P - \overline{Y}_1^P - \Delta$$

Thus, the difference-in-differences estimate of program impact is the change over time that participants experienced minus an estimate of the change participants would have experienced in the absence of the program. The parallel trend assumption allows us to form an estimate of that counterfactual change from the change over time that non-participants actually did experience.

In practice, perhaps the most popular method of performing difference-in-differences estimation is to estimate a regression along the lines of

$$Y_{it} = \beta_0 + \beta_1 \cdot P_i \cdot d_t + \beta_2 \cdot P_i + \beta_3 \cdot d_t + \epsilon_{it}$$

where d_t equals 1 if $t = 2$ and 0 if $t = 1$. P_i equals 1 if individual i is *ever* a participant in the program (which in the context of this example means if they participate at time $t = 2$, after the program is introduced) and 0 otherwise. ϵ_{it} is an error term, independently and identically distributed across individuals and within each individual across time. Sometimes additional time-varying controls are added to this specification in the hope that they will capture and control for any divergence from the parallel trend that would otherwise exist between participants and non-participants.

Given the discussion of sample size estimation for regression, it should not come as a big surprise to learn that direct estimation of appropriate sample size from this specification via the sample size estimation strategies outlined in the last section might be a challenge. The most elemental estimator of program impact via the difference-in-differences strategy would be

$$\left(\hat{\overline{Y}}_2^P - \hat{\overline{Y}}_1^P\right) - \left(\hat{\overline{Y}}_2^N - \hat{\overline{Y}}_1^N\right)$$

where

$$\hat{\overline{Y}}_t^j$$

is an estimate of the population outcome for population j (where j can equal P or N) at time t (where in our simple formulation thus far we have considered just 2 time periods, $t = 1, 2$). Various more involved extensions of this estimator have been proposed (for example, Heckman, Ichimura, and Todd [1997] and Heckman, Ichimura, Smith, and Todd [1998]) develop a matching version of this estimator) and it can be shown that this is what the term β_3 from the regression version of the difference-in-differences strategy estimates.

Though this is a simple estimator of impact via difference-in-differences and as such is probably not employed that often in current impact evaluations (where the "technology" seems to be drifting toward more elaborate regression, matching, "pre-matching," etc.) its sample size implications are quite interesting. To begin with, for parsimony define

$$I = \left(\overline{Y}_2^P - \overline{Y}_1^P\right) - \left(\overline{Y}_2^N - \overline{Y}_1^N\right)$$

I stands for "impact" as estimated by the difference-in-differences approach. Estimated impact is then

$$\hat{I} = \left(\hat{\overline{Y}}_2^P - \hat{\overline{Y}}_1^P\right) - \left(\hat{\overline{Y}}_2^N - \hat{\overline{Y}}_1^N\right)$$

Now, suppose that we wish to determine a sample size large enough to accommodate the null and alternative hypotheses

$$H_0 : I = I^*$$

$$H_a : I \neq I^*$$

with the specific alternative hypothesis

$$H_a : I \leq I^* - d \text{ or } I \geq I^* + d$$

We are thus setting ourselves up for a two-sided test[109] that the impact is (null hypothesis) some value I^* or (alternative hypothesis) differs from that value by a margin of d or more. For instance, if $I^* = 0$, then this is basically testing whether the program had no impact against an alternative hypothesis that it had an impact of d or more.

We accept (again, we say this instead of "fail to reject" simply for brevity's sake) the null hypothesis at the α level of significance if the value of the test statistic t satisfies

$$|t| \leq z_{1-\frac{\alpha}{2}}$$

Thus we retain the basic design from the benchmark framework for testing for differences in indicators.

Let us set ourselves up to be able to derive a general form for the power function and sample size estimate in the two-sided case. We begin with the test statistic under the null hypothesis

$$t = \frac{\hat{I} - I^*}{\sqrt{V\left(\hat{I}\right)}}$$

[109] The one-sided case involves, as in the first section of this chapter, using the critical value for significance of $z_{1-\alpha}$ rather than the two-sided critical value of $z_{1-\alpha/2}$.

The probability of accepting the null if one side of the two-sided alternative is true:

$$Pr\left(|t| \leq z_{1-\frac{\alpha}{2}} \,\middle|\, I = I^* + d\right)$$

$$= Pr\left(-z_{1-\frac{\alpha}{2}} \leq \frac{\hat{I} - I^*}{\sqrt{V\left(\hat{I}\right)}} \leq z_{1-\frac{\alpha}{2}} \,\middle|\, I = I^* + d\right)$$

This is *exactly* the same setup as in our discussion of two-sided tests in the first section. And it is going to lead to the same place. The power function is

$$Pr\left(|t| > z_{1-\frac{\alpha}{2}} \,\middle|\, I = I^* + d\right) = 1 - Pr\left(|t| \leq z_{1-\frac{\alpha}{2}} \,\middle|\, I = I^* + d\right)$$

$$= 1 - Pr\left(Z \leq z_{1-\frac{\alpha}{2}} - \frac{d}{\sqrt{V\left(\hat{I}\right)}} \,\middle|\, I = I^* + d\right)$$

$$= Pr\left(Z > z_{1-\frac{\alpha}{2}} - \frac{d}{\sqrt{V\left(\hat{I}\right)}} \,\middle|\, I = I^* + d\right)$$

If we want the sample size, we again set

$$z_{1-\frac{\alpha}{2}} - \frac{d}{\sqrt{V\left(\hat{I}\right)}} = -z_{1-\beta}$$

and solve for n.

What thus remains for deriving a sample size estimate is to determine a functional form for

$$V\left(\hat{I}\right) = V\left(\left(\hat{\overline{Y}}_2^P - \hat{\overline{Y}}_1^P\right) - \left(\hat{\overline{Y}}_2^N - \hat{\overline{Y}}_1^N\right)\right)$$

$$= V\left(\hat{\overline{Y}}_2^P - \hat{\overline{Y}}_1^P - \hat{\overline{Y}}_2^N + \hat{\overline{Y}}_1^N\right)$$

We next need to understand something about the variance of linear combinations of random variables. If $W_1, ..., W_R$ are R (possibly correlated) random variables and $c_1, .., c_R$ are a bunch of constants, then

$$Var\left(\sum_{r=1}^{R} c_r \cdot W_r\right) = \sum_{r=1}^{R} c_r^2 \cdot V\left(W_r\right) + 2 \cdot \left(\sum_{r=1}^{R}\sum_{q>r}^{R} c_r \cdot c_q \cdot Cov\left(W_r, W_q\right)\right)$$

Thus we have[110]

$$V\left(\hat{I}\right) = V\left(\hat{\bar{Y}}_2^P - \hat{\bar{Y}}_1^P - \hat{\bar{Y}}_2^N + \hat{\bar{Y}}_1^N\right) = var\left(\hat{\bar{Y}}_2^P\right) + var\left(\hat{\bar{Y}}_1^P\right) + var\left(\hat{\bar{Y}}_2^N\right) + var\left(\hat{\bar{Y}}_1^N\right)$$

$$+2\cdot\left[-cov\left(\hat{\bar{Y}}_2^P, \hat{\bar{Y}}_1^P\right) - cov\left(\hat{\bar{Y}}_2^P, \hat{\bar{Y}}_2^N\right) + cov\left(\hat{\bar{Y}}_2^P, \hat{\bar{Y}}_1^N\right)\right.$$

$$\left.+cov\left(\hat{\bar{Y}}_1^P, \hat{\bar{Y}}_2^N\right) - cov\left(\hat{\bar{Y}}_1^P, \hat{\bar{Y}}_1^N\right) - cov\left(\hat{\bar{Y}}_2^N, \hat{\bar{Y}}_1^N\right)\right]$$

This is a pretty monstrous expression but we can probably simplify it a lot through some reasonable assumptions.

First, let us assume that the variance is the same across the pre-/post-program and participant/non-participant populations:

$$var\left(\hat{\bar{Y}}_2^P\right) = var\left(\hat{\bar{Y}}_1^P\right) = var\left(\hat{\bar{Y}}_2^N\right) = var\left(\hat{\bar{Y}}_1^N\right) = var\left(\hat{\bar{Y}}\right)$$

Strictly speaking this is probably not a reasonable assumption (one is basically assuming, for instance, that neither time not participation has any effect on the variance, which seems a stretch). However, it is likely a practical one in the sense that in many circumstances we are likely to lack decent priors about how variance differs across time and/or participation status. Furthermore, since this common variance is still that of an estimator, we have

$$var\left(\hat{\bar{Y}}\right) = \frac{\sigma_Y^2}{n}$$

This is the oft-invoked variance of an estimate of an average from the first section.

With this simplification in have we have

$$V\left(\hat{I}\right) = V\left(\hat{\bar{Y}}_2^P - \hat{\bar{Y}}_1^P - \hat{\bar{Y}}_2^N + \hat{\bar{Y}}_1^N\right) = \frac{\sigma_Y^2}{n} + \frac{\sigma_Y^2}{n} + \frac{\sigma_Y^2}{n} + \frac{\sigma_Y^2}{n}$$

$$+2\cdot\left[-cov\left(\hat{\bar{Y}}_2^P, \hat{\bar{Y}}_1^P\right) - cov\left(\hat{\bar{Y}}_2^P, \hat{\bar{Y}}_2^N\right) + cov\left(\hat{\bar{Y}}_2^P, \hat{\bar{Y}}_1^N\right)\right.$$

$$\left.+cov\left(\hat{\bar{Y}}_1^P, \hat{\bar{Y}}_2^N\right) - cov\left(\hat{\bar{Y}}_1^P, \hat{\bar{Y}}_1^N\right) - cov\left(\hat{\bar{Y}}_2^N, \hat{\bar{Y}}_1^N\right)\right]$$

$$V\left(\hat{I}\right) = V\left(\hat{\bar{Y}}_2^P - \hat{\bar{Y}}_1^P - \hat{\bar{Y}}_2^N + \hat{\bar{Y}}_1^N\right) = 4\cdot\frac{\sigma_Y^2}{n} + 2\cdot\left[-cov\left(\hat{\bar{Y}}_2^P, \hat{\bar{Y}}_1^P\right) - cov\left(\hat{\bar{Y}}_2^P, \hat{\bar{Y}}_2^N\right) + cov\left(\hat{\bar{Y}}_2^P, \hat{\bar{Y}}_1^N\right)\right.$$

$$\left.+cov\left(\hat{\bar{Y}}_1^P, \hat{\bar{Y}}_2^N\right) - cov\left(\hat{\bar{Y}}_1^P, \hat{\bar{Y}}_1^N\right) - cov\left(\hat{\bar{Y}}_2^N, \hat{\bar{Y}}_1^N\right)\right]$$

This is a somewhat "cleaned up" expression, but there is clearly potential scope for further tightening through the messy covariance terms.

Next, let us assume that the participant and non-participant groups are independent (in the

[110] Note that the constant term is 1 for $\hat{\bar{Y}}_2^P$ and $\hat{\bar{Y}}_1^N$ and -1 for $\hat{\bar{Y}}_1^P$ and $\hat{\bar{Y}}_2^N$ here.

sense that there is not correlation between the estimates from them). The more independent is the selection of the samples of participants and non-participants, the more reasonable is this assumption. For instance, it would likely not be reasonable if participants and non-participants were drawn from the same clusters.

With this assumption in hand we have

$$cov\left(\hat{\overline{Y}}_1^P, \hat{\overline{Y}}_2^N\right) = cov\left(\hat{\overline{Y}}_1^P, \hat{\overline{Y}}_1^N\right) = cov\left(\hat{\overline{Y}}_2^P, \hat{\overline{Y}}_2^N\right) = cov\left(\hat{\overline{Y}}_2^P, \hat{\overline{Y}}_1^N\right) = 0$$

Our expression thus further reduces to

$$\begin{aligned}
V\left(\hat{I}\right) = V\left(\hat{\overline{Y}}_2^P - \hat{\overline{Y}}_1^P - \hat{\overline{Y}}_2^N + \hat{\overline{Y}}_1^N\right) &= 4 \cdot \frac{\sigma_Y^2}{n} + 2 \cdot \left[-cov\left(\hat{\overline{Y}}_1^P, \hat{\overline{Y}}_2^P\right) - cov\left(\hat{\overline{Y}}_2^N, \hat{\overline{Y}}_1^N\right)\right] \\
&= 4 \cdot \frac{\sigma_Y^2}{n} - 2 \cdot \left[cov\left(\hat{\overline{Y}}_1^P, \hat{\overline{Y}}_2^P\right) + cov\left(\hat{\overline{Y}}_2^N, \hat{\overline{Y}}_1^N\right)\right] \\
&= 4 \cdot \frac{\sigma_Y^2}{n} - 2 \cdot \frac{\sigma_Y^2}{n} \cdot \left(\frac{cov\left(\hat{\overline{Y}}_1^P, \hat{\overline{Y}}_2^P\right)}{\frac{\sigma_Y^2}{n}} + \frac{cov\left(\hat{\overline{Y}}_2^N, \hat{\overline{Y}}_1^N\right)}{\frac{\sigma_Y^2}{n}}\right)
\end{aligned}$$

For those who read through section 1 thoroughly, their should be an emerging familiarity to this.

Now, by one of our earlier simplifying assumptions, we have[111]

$$cov\left(\hat{\overline{Y}}_1^P, \hat{\overline{Y}}_2^P\right) = \frac{1}{n} \cdot \rho \cdot \sigma_{Y_1^P} \cdot \sigma_{Y_2^P} = \frac{1}{n} \cdot \rho\sigma_Y^2$$

By similar logic, we have

$$cov\left(\hat{\overline{Y}}_1^N, \hat{\overline{Y}}_2^N\right) = \frac{1}{n} \cdot \rho\sigma_Y^2$$

Then we have

$$\begin{aligned}
V\left(\hat{I}\right) = V\left(\hat{\overline{Y}}_2^P - \hat{\overline{Y}}_1^P - \hat{\overline{Y}}_2^N + \hat{\overline{Y}}_1^N\right) &= 4 \cdot \frac{\sigma_Y^2}{n} - 2 \cdot \frac{\sigma_Y^2}{n} \cdot \left(\frac{cov\left(\hat{\overline{Y}}_1^P, \hat{\overline{Y}}_2^P\right)}{\frac{\sigma_Y^2}{n}} + \frac{cov\left(\hat{\overline{Y}}_2^N, \hat{\overline{Y}}_1^N\right)}{\frac{\sigma_Y^2}{n}}\right) \\
&= 4 \cdot \frac{\sigma_Y^2}{n} - 2 \cdot \frac{\sigma_Y^2}{n} \cdot \left(\frac{\frac{1}{n} \cdot \rho\sigma_Y^2}{\frac{\sigma_Y^2}{n}} + \frac{\frac{1}{n} \cdot \rho\sigma_Y^2}{\frac{\sigma_Y^2}{n}}\right) \\
&= 4 \cdot \frac{\sigma_Y^2}{n} - 2 \cdot \frac{\sigma_Y^2}{n} \cdot (\rho + \rho) \\
&= 4 \cdot \frac{\sigma_Y^2}{n} - 4 \cdot \frac{\sigma_Y^2}{n} \cdot \rho
\end{aligned}$$

[111] We have actually snuck in the additional assumption that the intertemporal correlation of the estimate of the mean is ρ for participants and non-participants.

$$= 4 \cdot \frac{\sigma_Y^2}{n} \cdot (1-\rho)$$

We now have a variance formula that we can work with for deriving sample size.[112]

Returning to the condition

$$z_{1-\frac{\alpha}{2}} - \frac{d}{\sqrt{V\left(\hat{I}\right)}} = -z_{1-\beta}$$

We then have

$$z_{1-\frac{\alpha}{2}} - \frac{d}{\sqrt{4 \cdot \frac{\sigma_Y^2}{n} \cdot (1-\rho)}} = -z_{1-\beta}$$

or

$$z_{1-\frac{\alpha}{2}} + z_{1-\beta} = \frac{d}{\sqrt{4 \cdot \frac{\sigma_Y^2}{n} \cdot (1-\rho)}}$$

leading to

$$\sqrt{\frac{4 \cdot \sigma^2 \cdot (1-\rho)}{n}} = \frac{d}{z_{1-\frac{\alpha}{2}} + z_{1-\beta}}$$

and then

$$\frac{1}{\sqrt{n}} = \frac{d}{\sqrt{4 \cdot \sigma^2 \cdot (1-\rho)} \cdot \left(z_{1-\frac{\alpha}{2}} + z_{1-\beta}\right)}$$

Inverting both sides and squaring, we have

$$n = 4 \cdot (1-\rho) \left(\left[\frac{\sigma \cdot \left(z_{1-\frac{\alpha}{2}} + z_{1-\beta}\right)}{d} \right]^2 \right)$$

Once again, this would be the sample size for *each* sample (participants before program implementation, non-participants before program implementation, participants after program implementation and non-participants after program implementation).

This is a fairly breathtaking finding. It suggests the possibility of enormous required sample sizes for difference-in-differences applications. Of course, to the extent that there is a correlation over time ρ we should be able to reduce the estimate of the required sample size.

The trouble is that appeal to ρ is, particularly if we plan to go the regression estimation route, useful only if one can then properly model its role in the error terms of the difference-in-differences model. For a long time, difference-in-differences proved to be an attractive model for practitioners, and for good reason: it seemed to offer (within and well beyond the program impact evaluation literature) a seemingly unending series of statistically significant estimates of causal links between all kinds of variables, often with quite modest sample sizes. Then came along a key paper that somewhat upset this apple cart: Bertrand, Duflo, and Mullainathan (2004) found that, in the

[112] Though we derived this variance ourselves, to our knowledge it was also independently derived by Anthony Turner and, to our knowledge, more or less by Frison and Pocock (1993). Indeed, Frison and Pocock derive for the more general case (with potentially multiple time periods observed before and after program implementation or commencement). See McKenzie (2012) for an interesting application of Frison and Pocock's basic results.

presence of serially correlated outcomes, the model found significant relationships between variables where none actually existed far more often than conventional assumptions of Type-I error rates would predict.[113] What Bertrand, et al. suggested is that, in the presence of serial correlation, the difference-in-differences model loads the dice, to a surprisingly great deal, toward finding significant causal relationships between the variables under study. It thus becomes difficult to know how much significant findings can be trusted when difference-in-differences is applied to potentially serially correlated data. Bertrand, et al. try various remedies for this, with generally limited success.

In general, the problem that Bertrand, et al. explore appears to be one largely of group mean models (Angrist & Pischke, 2009). The difference-in-differences model is a group mean one in the sense that, for instance, we estimate overall means for groups (participants and non-participants at different time periods) as opposed to individual-level fixed effects. One remedy the authors have used and have seen used by others for the issues revealed by Bertrand, et al. (2004) is the use of the dummy variable fixed effects estimator, with a dummy for each individual participant and non-participant.[114] While this solution may help some with serial correlation of errors, it is an expensive one in terms of additional parameters that need to be estimated.

We turn briefly to sample size estimation for the difference-in-differences model when the outcome is a proportion. Again, our focus is on the simple implementation

$$I = \left(\overline{Y}_2^P - \overline{Y}_1^P\right) - \left(\overline{Y}_2^N - \overline{Y}_1^N\right)$$

with the corresponding estimator

$$\hat{I} = \left(\hat{\overline{Y}}_2^P - \hat{\overline{Y}}_1^P\right) - \left(\hat{\overline{Y}}_2^N - \hat{\overline{Y}}_1^N\right)$$

To simplify the discussion, we assume that the samples sizes are the same across the four "domains" (participants before program implementation, participants after program implementation, non-participants before program implementation, and non-participants after program implementation). We also assume that the estimates are independent across time (i.e., we assume that $\rho = 0$). Finally, we assume that $I^* = 0$ (which is quite reasonable, as this is typically how the null hypothesis would be framed).

The variance estimates for

$$\left(\overline{Y}_2^P - \overline{Y}_1^P\right) - \left(\overline{Y}_2^N - \overline{Y}_1^N\right) = 0$$

are a little trickier now than in the case of the simple comparison of percentages or rates between populations because we cannot simply assume that all of the proportions will be the same under the null hypothesis: $\overline{Y}_2^N$ and $\overline{Y}_1^N$ will differ to the extent that there is a trend to outcomes over

[113] Specifically, with a panel data set of women from the United States, Bertrand, et al. (2004) simply made up labor market policies and then found enormous Type-I error rates for regression difference-in-differences estimates of the impact of those fake policies on labor market outcomes.

[114] Technically, at least one must be withheld for identification purposes.

time. The same will be the case with $\overline{Y}_2^P$ and $\overline{Y}_1^P$. Indeed, if the parallel trend assumption holds, the program has no impact and the parallel trend is ξ, then we have

$$\overline{Y}_2^N - \overline{Y}_1^N = \overline{Y}_2^P - \overline{Y}_1^P = \xi$$

and

$$\left(\overline{Y}_2^P - \overline{Y}_1^P\right) - \left(\overline{Y}_2^N - \overline{Y}_1^N\right) = 0$$

The fact that the various averages don't simply have the same value (because $\overline{Y}_2^P \neq \overline{Y}_1^P$ and $\overline{Y}_2^N \neq \overline{Y}_1^N$) makes it difficult to justify use of the "pooled estimator" mentioned earlier.

There are two options in this case, which we will call "option A" and "option B." Option A involves assuming that ξ is very small. This would imply that, under the null hypothesis

$$\overline{Y}_2^N - \overline{Y}_1^N \approx 0$$

and

$$\overline{Y}_2^P - \overline{Y}_1^P \approx 0$$

Then we could in some sense rationalize the use of the pooled variance estimator, leaving us with

$$V\left(\overline{Y}_2^P - \overline{Y}_1^P - \overline{Y}_2^N + \overline{Y}_1^N\right) \approx 4 \cdot \frac{\overline{Y} \cdot \left(1 - \overline{Y}\right)}{n}$$

where

$$\overline{Y} = \frac{\overline{Y}_2^P + \overline{Y}_1^P + \overline{Y}_2^N + \overline{Y}_1^N}{4}$$

This is sort of a sketchy option since, by using a tiny ξ as its departure point, it more or less rules out the possibility of a secular trend to outcomes in the absence of program participation. If this was the case one would not really need to use a difference-in-differences design and could instead rely on a simpler design (the basic pre-/post-design comes to mind).

Option B involves admitting that there is a secular (i.e., in the absence of the program) time trend to outcomes (in other words, we assume that ξ has teeth) and, therefore,

$$\overline{Y}_2^N \neq \overline{Y}_1^N$$

and

$$\overline{Y}_2^P \neq \overline{Y}_1^P$$

Let us define

$$\tilde{\overline{Y}}_2^N = \overline{Y}_1^N + \xi$$

and

$$\tilde{\overline{Y}}_2^P = \overline{Y}_1^P + \xi$$

Then one estimate of the variance would be

$$V\left(\overline{Y}_2^P - \overline{Y}_1^P - \overline{Y}_2^N + \overline{Y}_1^N\right)$$

$$= \frac{\tilde{\overline{Y}}_2^P \cdot \left(1 - \tilde{\overline{Y}}_2^P\right) + \overline{Y}_1^P \cdot \left(1 - \overline{Y}_1^P\right) + \tilde{\overline{Y}}_2^N \cdot \left(1 - \tilde{\overline{Y}}_2^N\right) + \overline{Y}_1^N \cdot \left(1 - \overline{Y}_1^N\right)}{n}$$

Implementing this variance estimate would require more prior assumptions, but it has the advantage of not relying on an assumption of no secular trend to outcomes.

The variance under the alternative is more analytically clear. Define

$$\breve{\overline{Y}}_2^P = \overline{Y}_1^P + \xi + d$$

The variance is then

$$V\left(\overline{Y}_2^P - \overline{Y}_1^P - \overline{Y}_2^N + \overline{Y}_1^N\right)$$

$$= \frac{\breve{\overline{Y}}_2^P \cdot \left(1 - \breve{\overline{Y}}_2^P\right) + \overline{Y}_1^P \cdot \left(1 - \overline{Y}_1^P\right) + \tilde{\overline{Y}}_2^N \cdot \left(1 - \tilde{\overline{Y}}_2^N\right) + \overline{Y}_1^N \cdot \left(1 - \overline{Y}_1^N\right)}{n}$$

This is simply the straightforward extension of the logic of the variance under the alternative from the two-sample case from the first section of this chapter.

To simplify the discussion that follows, let us define the following:

$$V_0^{\text{Option 1}} = 4 \cdot \overline{Y} \cdot \left(1 - \overline{Y}\right)$$

$$V_0^{\text{Option 2}} = \tilde{\overline{Y}}_2^P \cdot \left(1 - \tilde{\overline{Y}}_2^P\right) + \overline{Y}_1^P \cdot \left(1 - \overline{Y}_1^P\right) + \tilde{\overline{Y}}_2^N \cdot \left(1 - \tilde{\overline{Y}}_2^N\right) + \overline{Y}_1^N \cdot \left(1 - \overline{Y}_1^N\right)$$

$$V_1 = \breve{\overline{Y}}_2^P \cdot \left(1 - \breve{\overline{Y}}_2^P\right) + \overline{Y}_1^P \cdot \left(1 - \overline{Y}_1^P\right) + \tilde{\overline{Y}}_2^N \cdot \left(1 - \tilde{\overline{Y}}_2^N\right) + \overline{Y}_1^N \cdot \left(1 - \overline{Y}_1^N\right)$$

These are, respectively, the numerator of the variance terms (the denominator being n in all cases) under the null hypothesis Option 1 variance, under the null hypothesis Option 2 variance, and under the alternative hypothesis.

Under Option 1 sample size is then

$$n = \left[\frac{z_{1-\alpha/2} \cdot \sqrt{V_0^{\text{Option 1}}} + z_{1-\beta} \cdot \sqrt{V_1}}{d}\right]^2$$

while under Option 2 it is

$$n = \left[\frac{z_{1-\alpha/2} \cdot \sqrt{V_0^{\text{Option 2}}} + z_{1-\beta} \cdot \sqrt{V_1}}{d}\right]^2$$

In either case, it is useful to compare these formulas to the sample size formulas for comparison of means between two populations case discussed in the first section of this chapter:

$$n = \left[\frac{z_{1-\alpha/2} \cdot \sqrt{2 \cdot \overline{\overline{Y}} \cdot \left(1 - \overline{\overline{Y}}\right)} + z_{1-\beta} \cdot \sqrt{\overline{Y}_1 \cdot \left(1 - \overline{Y}_1\right) + \overline{Y}_2 \cdot \left(1 - \overline{Y}_2\right)}}{d} \right]^2$$

Under the Option 1 and Option 2 formulas, n is going to be larger, in many cases much larger. Clearly, the dramatic increase in sample size (over the simple comparison of means case) for the continuous case carries over to proportions.[115]

This has been a bit of a cautionary tale in terms of sample sizes for difference-in-differences. The sample sizes for this, the simplest implementation of difference-in-differences, are going to be comparatively big.

3.5 Instrumental Variables

3.5.1 The Basics

We now turn to the final common impact evaluation strategy: instrumental variables. The determination of the potential outcomes for individual i is as follows:

$$Y_i^0 = \beta_0 + \beta_2 \cdot X_i + \beta_3 \cdot \mu_i + \epsilon_i^Y$$

$$Y_i^1 = \beta_0 + \beta_1 + \beta_2 \cdot X_i + \beta_3 \cdot \mu_i + \epsilon_i^Y$$

We have an observed (X) and unobserved (μ) individual-level characteristic determining potential outcomes as well as a purely random, idiosyncratic unobserved component ϵ^Y. We assume that the three are independently distributed. Program impact for individual i is

$$\begin{aligned} & Y_i^1 - Y_i^0 \\ &= \beta_0 + \beta_1 + \beta_2 \cdot X_i + \beta_3 \cdot \mu_i + \epsilon_i^Y \\ &- \left(\beta_0 + \beta_2 \cdot X_i + \beta_3 \cdot \mu_i + \epsilon_i^Y\right) \\ &= \beta_1 \end{aligned}$$

Thus, program impact is constant across individuals.

Note that we have dropped the time subscript t in these potential outcome equations. We will do the same in the cost equation to be introduced shortly. We no longer distinguish between fixed and time-varying observed or unobserved characteristics. Although instrumental variables

[115] One stone we have left unturned is what to do about the possibility of intertemporal correlation as captured by ρ in the continuous case. As in the first section discussion of comparison of proportions between two populations, we would tend to use the "cheat" of multiplying the sample size estimate n by $(1 - \rho)$.

is sometimes applied in the panel data setting, it is a cross-sectional method in the sense that within variation or even repeated measures *per se* are not intrinsically necessary to apply it. The logic of identification of program impact (i.e., the manner in which instrumental variables allows for consistent estimation of program impact when program participation is endogenous) that we establish in the cross-sectional case applies to any other setting where we might apply it (such as to the change in program participation in a panel data setting where one might use both the within and instrumental variables approaches simultaneously).[116]

A regression specification can be derived in much the same fashion as with behavioral models in preceding sections. The observed outcome is

$$\begin{aligned} Y_i &= P_i \cdot Y_i^1 + (1 - P_i) \cdot Y_i^0 \\ &= P_i \cdot \left(\beta_0 + \beta_1 + \beta_2 \cdot X_i + \beta_3 \cdot \mu_i + \epsilon_i^Y\right) \\ &+ (1 - P_i) \cdot \left(\beta_0 + \beta_2 \cdot X_i + \beta_3 \cdot \mu_i + \epsilon_i^Y\right) \\ &= \beta_0 + \beta_1 \cdot P_i + \beta_2 \cdot X_i + \beta_3 \cdot \mu_i + \epsilon_i^Y \end{aligned}$$

Notice that the unobservable μ_i appears in this equation, and would hence be an element of the regression residual.

The cost of program participation is given by

$$C_i = \gamma_0 + \gamma_1 \cdot X_i + \gamma_2 \cdot Z_i + \gamma_3 \cdot \mu_i + \epsilon_i^C$$

With one exception (the variable Z) this is, given the potential outcome equations, a fairly standard cost function in the tradition of the behavioral models discussed thus far. Individual i will choose to participate (i.e., their value for the program participation indicator P, P_i, equals 1) if

$$Y_i^1 - Y_i^0 - C_i > 0$$

or

$$\beta_1 - \gamma_0 - \gamma_1 \cdot X_i - \gamma_2 \cdot Z_i - \gamma_3 \cdot \mu_i - \epsilon_i^C > 0$$

where we assume that ϵ_i^C is uncorrelated with both ϵ_i^Y and the variables Z_i.

There are several important things to note about this. First, the observed individual characteristic X_i influences participation and hence we would expect that, other things being equal, the average value of this characteristic should differ between participants and non-participants. There will thus clearly be selection into program participation based on observables. Second, the same is true of the unobserved characteristic μ_i. Finally, we have an individual characteristic Z_i that does not influence directly the outcome of interest Y_i. Z_i does influence the program participation

[116] The application of the instrumental variables approach in the context of a within model would most likely be motivated by concerns of confounding *time-varying* unobserved characteristics, since the within model itself would address any concerning *fixed* unobserved characteristics.

decision, and is independent of either of the unobserved determinants of the outcome, μ and ϵ_i^Y (we introduce no avenue for there to be correlation). In other words, Z meets all of the criteria for an instrument.

The usual estimation strategy in this case is referred to as "two-stage least squares." The strategy is typically implemented as follows:

1. By regressing P on Z and X, estimate by ordinary least squares regression the model:

$$P_i = \kappa_1 + \kappa_2 \cdot Z_i + \kappa_3 \cdot X_i + \upsilon_i$$

2. Compute predicted program participation from the fitted model:

$$\hat{P}_i = \hat{\kappa}_1 + \hat{\kappa}_2 \cdot Z_i + \hat{\kappa}_3 \cdot X_i$$

3. Regress Y_i on $\hat{P}_i$ and X_i.

This is an attractive strategy when program participation is non-random and one has an instrument available.

The reader would be reasonable to ask about the behavioral basis for the equation

$$P_i = \kappa_1 + \kappa_2 \cdot Z_i + \kappa_3 \cdot X_i + \upsilon_i$$

This specification is easily motivated by the program participation condition

$$Y_i^1 - Y_i^0 - C_i > 0$$

or

$$\beta_1 - \gamma_0 - \gamma_1 \cdot X_i - \gamma_2 \cdot Z_i - \gamma_3 \cdot \mu_i - \epsilon_i^C > 0$$

This condition determines the value of the binary condition of program participation captured by the variable P_i. This condition tells us that P_i will depend on the observed variables X_i and Z_i and the unobservables μ_i and ϵ_i^C. The model

$$P_i = \kappa_1 + \kappa_2 \cdot Z_i + \kappa_3 \cdot X_i + \upsilon_i$$

can be thought of as a linear approximation to this condition. Essentially, it is a linear probability model of P_i.

Two-stage least squares is typically framed in terms of continuous first and second stage outcomes. Certainly that is the case that motivates most textbook introductions to it. Both stages are then typically estimated via linear regression (i.e., OLS).[117] In our case, the first stage outcome

[117] It actually isn't quite that simple, since separate, independent OLS estimation of the first and second stage model will yield incorrect second-stage standard errors. The reason for this is that straightforward, conventional estimates of the standard errors under OLS do not account for the sampling variation associated with the fact that $\hat{P}_i$ is an estimate generated by another OLS regression.

program participation P is usually not continuous[118] but instead modeled as

$$P_i = \begin{cases} 1 & \text{if individual } i \text{ participates in the program} \\ 0 & \text{if individual } i \text{ does not to participate in the program} \end{cases}$$

Hence P_i is a binary variable. Moreover, we will sometimes be interested in cases where the outcome of interest Y is a binary variable. For example, we might be interested in the impact of participation in a program on the probability of using modern contraception. There might be some temptation to use an explicit limited dependent variable model (such as logit or probit) when the outcome is binary. However, this is not necessary and might even be harmful (Angrist & Krueger, 2001). It is perfectly legitimate to estimate the two-stage least squares estimator via OLS at each stage even when either or both of the outcomes is binary. Moreover, there is evidence (such as Guilkey & Lance, 2014) that this estimation strategy often performs comparatively well.

Assuming the instrument meets the necessary criteria described for an instrument (namely, that the instrument is not correlated with ϵ_i^Y and influences Y only through P),[119] two-stage least squares yields an estimate $\hat{\beta}_1$ of β_1 that is consistent. A estimator is consistent if the distribution of the possible values it can take on becomes increasingly concentrated on the true population value (in this case true program impact β_1) as the sample size n becomes increasingly large. This suggests that, in general, larger sample sizes (compared with, for instance, those required to detect a given program effect with a given significance and power by simple comparison of sample estimates of mean outcomes between samples of participants and non-participants) are likely safer where instrumental variables are concerned (such as Angrist & Krueger, 2001; Boef, Dekkers, Vandenbroucke, & le Cessie, 2014). What "large" means is, of course, difficult to define precisely simply from the observation that instrumental variables is a consistent (as opposed to unbiased) estimator of program impact β_1.

Aside from the intrinsic consistency of the instrumental variables two-stage least squares estimator, another challenge identified more recently is the weak instrument problem (such as Bound, Jaeger, & Baker, 1995; Staiger & Stock, 1997). Put simply, when Z is only a weak predictor of P, the two-stage least squares estimator performs quite poorly, generating estimates with mean squared errors that suggest one might actually do better simply by regressing Y on P and Z. This problem can be eased to a degree with larger sample sizes (at which it might be possible to model accurately a slight relationship between P and Z). However, Bound, et al. (1995), Staiger and Stock (1997) and other work in this area suggests that larger sample size is not guaranteed to eliminate the problem. Indeed, a numerical example from the discussion of instrumental variables in Lance, Guilkey, et al. (2014) illustrates the potential persistence of this problem, even when there is some intrinsic relationship between program participation P and the instrument Z and sample size is enormous (at least by the normal standards of empirical practice that we have seen).

Even when there is a relationship between program participation P and the instrument Z, one

[118] It can sometimes take on a continuous or pseudo-continuous character when one considers a dose-response angle, for instance.

[119] And, throughout this discussion, we make this assumption.

might still need a reasonably large sample size to get a pulse on that first-stage relationship in the form of a significant, reasonably precise estimate of κ_2. Where all of these roads are leading is that instrumental variables estimation is typically a data-hungry (in the sense of requiring a healthy number observations n to achieve significance and power targets given the alternative hypothesis) impact estimation strategy.

3.5.2 Sample Size Estimation

In light of all of the limitations, complexity, shadows and fog surrounding sample size estimation for selection on observables models (and hence, by extension, within models and to a certain extent RCTs), there is a surprisingly and refreshingly straightforward basic but useful sample size estimation formula for instrumental variables. The crucible for development of this sample size estimation approach has been the "Mendelian randomization" literature.

"Mendelian" is a reference to Gregor Mendel (1822 – 1884), the Moravian scientist and friar who was essentially the founder of the science of genetics. Mendelian randomization refers to the use of genetic markers as instruments for risk factors for the purpose of identifying the impact of those risk factors on health. So, to make explicit how this would fit into the structure of our program impact challenge, we have:

1. Z_i: the genetic marker
2. P_i: the risk factor
3. Y_i the health outcome

Although the sample size estimation method we will discuss was motivated by a completely different goal (to identify the impact of risk factors on outcomes like health), the basic structure of the problem is analogous. Hence, the sample size estimation developed for Mendelian randomization is quite applicable to the impact evaluation context.

Our discussion of this will integrate elements of numerous manuscripts (Freeman, Cowling, & Schooling, 2013; Burgess, 2014). We begin with the "Wald estimator" (we put it in quotation marks because this is how the Wald estimator is usually presented in this literature, but it has another common guise in the sample size estimation literature rooted in Mendellian randomization; more on this below) commonly associated with two-stage least squares instrumental variables estimation. The Wald estimator is an important way of characterizing the two-stage least squares estimator.[120] The Wald estimator is

$$\hat{\beta}_1^{IV} = \frac{cov\left(Y, Z\right)}{cov\left(P, Z\right)}$$

The numerator is thus the covariance between the outcome Y and the instrument Z divided by the covariance between program participation P and the instrument Z.

[120] See Angrist and Pischke (2009) for a very thorough discussion of two-stage least squares and instrumental variables more generally that revolves to a significant degree around the Wald estimator.

This estimator can be contrasted with the ordinary least squares estimator of program impact. Making that contrast is not that easy under the current model (whereby Y is regressed on P and X to obtain an estimate of β_1) because the linear regression model with more than one regressor is usually expressed in terms of matrix algebra (and, as daunting as matrix algebra sounds, the formula for the linear regression estimation of β_1 is actually more visually challenging without it).

The easy way around this for the purposes of this discussion is to reduce it back to a one regressor model. This is easily done. Simply assume that $\beta_2 = \gamma_1 = 0$. The ordinary least squares estimator of β_1, $\hat{\beta}_1^{OLS}$ is

$$\hat{\beta}_1^{OLS} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\cdot\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} = \frac{\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\cdot\left(Y_i - \overline{Y}\right)}{n}}{\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n}} = \frac{cov\left(X,Y\right)}{var\left(X\right)}$$

There are a number of ways of deriving the Wald estimator for this single instrument case. First, in the interest of fair disclosure, the "Wald estimator" in the context of two-stage least squares with a single instrument is often presented as

$$\hat{\beta}_1^{IV} = \frac{E\left[Y|Z=1\right] - E\left[Y|Z=0\right]}{E\left[P|Z=1\right] - E\left[P|Z=0\right]}$$

While this might seem completely different from the "Wald estimator"

$$\hat{\beta}_1^{IV} = \frac{cov\left(Y,Z\right)}{cov\left(P,Z\right)}$$

in actuality they are the same thing.

To begin with, the basic logic of instrumental variables is to use an instrument ("Z") to find a channel of variation in an explanatory variable P that is "random" in the sense that it is uncorrelated with any unobserved factors that determine Y (unobserved determinants of Y might also be associated with X, making it impossible to estimate the causal impact of P on Y using the full variation observed in P). The first stage in two-stage least squares regression, involving regression of P on Z and then prediction of P from the fitted model thus separates the variation in P into two channels: "experimental" variation in P (associated with variation in Z) and everything else. For the purpose of estimating the causal impact of P on Y in the second stage, we retain only the former. The variation in Z is correlated with, and hence traces out (or, more formally and appropriately to the language of statistics, "identifies") random or experimental variation in P.

It is therefore reasonable to think of this estimator in terms of the impact of Z on Y. This impact emerges due to the impact of Z on P. In other words,

$$\begin{pmatrix}\text{Impact of } Z\\ \text{on } Y\end{pmatrix} = \begin{pmatrix}\text{Impact of } P\\ \text{on } Y\end{pmatrix}\cdot\begin{pmatrix}\text{Impact of } Z\\ \text{on } P\end{pmatrix}$$

However, the true impact of P on Y is β_1. Thus, we have

$$\begin{pmatrix} \text{Impact of } Z \\ \text{on } Y \end{pmatrix} = \beta_1 \cdot \begin{pmatrix} \text{Impact of } Z \\ \text{on } P \end{pmatrix}$$

or, solving for β_1,

$$\beta_1 = \frac{\begin{pmatrix} \text{Impact of } Z \\ \text{on } Y \end{pmatrix}}{\begin{pmatrix} \text{Impact of } Z \\ \text{on } P \end{pmatrix}}$$

All of the versions of the "Wald" estimator reflect this logic.

First, consider

$$\hat{\beta}_1^{IV} = \frac{cov\left(Y,Z\right)}{cov\left(P,Z\right)}$$

Dividing the numerator and denominator by $var(Z)$ we have

$$\hat{\beta}_1^{IV} = \frac{\frac{cov(Y,Z)}{var(Z)}}{\frac{cov(P,Z)}{var(Z)}}$$

However,

$$\frac{cov\left(Y,Z\right)}{var\left(Z\right)}$$

is the linear regression estimator (i.e., estimator from regressing Y on Z) of the impact of Z on Y while

$$\frac{cov\left(P,Z\right)}{var\left(Z\right)}$$

is the same for the impact of Z on P. However, in the case where Z is binary, what regressing Y on Z really estimates is

$$E\left(Y|Z=1\right) - E\left(Y|Z=0\right)$$

In other words, we estimate the expected change in Y from switching Z from 0 to 1. Similarly, when Z is binary, regressing P on Z really estimates

$$E\left(P|Z=1\right) - E\left(P|Z=0\right)$$

Hence, the two formulations of the Wald estimator are essentially the same.

Because it is at the center of the discussion of sample size estimation for instrumental variables, the representation of the Wald estimator on which we will focus is

$$\hat{\beta}_1^{IV} = \frac{cov\left(Y,Z\right)}{cov\left(P,Z\right)}$$

When we say this, we refer to one particularly seminal paper, Freeman, et al. (2013). While there are other important papers in this literature, in what follows we derive the sample size estimator for the single instrument case in such a fashion that the derivation can be considered a quote from Freeman, et al.. Their presentation is so focused, elegant and parsimonious that it is difficult to paraphrase it in a fashion completely intellectually appropriate to the notion of paraphrasing.

To begin with, the Wald estimator is approximately distributed as

$$\hat{\beta}_1^{IV} \sim N\left(\beta_1, \frac{var\left(Y|P\right)}{n \cdot var\left(P\right) \cdot \rho_{PZ}^2}\right)$$

We emphasize that this is an approximation, and a more confidence-inspiring one the larger is n.[121] $var(Y|P)$ is the variance of Y not driven by P while ρ_{PZ}^2 is the correlation between P and Z. The former can be estimated as the variance of $Y - \hat{\beta}_1 \cdot P$ while the latter can be estimated as the R-squared for a regression of P on Z.

Anyway, under the null hypothesis

$$H_0 : \beta_1 = 0$$

of no program impact, the test statistic

$$\frac{\hat{\beta}_1^{IV}}{\sqrt{\frac{var(Y|P)}{n \cdot var(P) \cdot \rho_{PZ}^2}}}$$

has a (approximately) normal distribution. To avoid visually intense notation, we follow Freeman, et al. (2013) by writing

$$\sqrt{V} = \sqrt{\frac{var\left(Y|P\right)}{var\left(P\right)}}$$

leaving us with the test statistic

$$\frac{\hat{\beta}_1^{IV}}{\frac{\sqrt{V}}{\sqrt{n} \cdot \rho_{PZ}}}$$

We reject the null hypothesis if the test statistic is either larger than $z_{1-\frac{\alpha}{2}}$ or less than $-z_{1-\frac{\alpha}{2}}$ or if

$$\frac{\left|\hat{\beta}_1^{IV}\right|}{\frac{\sqrt{V}}{\sqrt{n} \cdot \rho_{PZ}}} > z_{1-\frac{\alpha}{2}}$$

or, rearranging,

$$\left|\hat{\beta}_1^{IV}\right| > z_{1-\frac{\alpha}{2}} \cdot \frac{\sqrt{V}}{\sqrt{n} \cdot \rho_{PZ}}$$

Given the variance from the assumed approximate distribution for the instrumental variables estimator, this is fairly straightforward.

[121] This should not be seen as a reasonable approximation of $\hat{\beta}_1^{IV}$ for smaller n. The performance of the estimator at very small n can be simply awful and completely unpredictable and defies (to our knowledge) straightforward analytical characterization.

Then, following Freeman, et al., if the true program impact is d, the probability of rejecting H_0 is

$$Pr\left(\left|\hat{\beta}_1^{IV}\right| > z_{1-\frac{\alpha}{2}} \cdot \frac{\sqrt{V}}{\sqrt{n} \cdot \rho_{PZ}}\right) = Pr\left(\hat{\beta}_1^{IV} < -z_{1-\frac{\alpha}{2}} \cdot \frac{\sqrt{V}}{\sqrt{n} \cdot \rho_{PZ}}\right)$$

$$+Pr\left(\hat{\beta}_1^{IV} > z_{1-\frac{\alpha}{2}} \cdot \frac{\sqrt{V}}{\sqrt{n} \cdot \rho_{PZ}}\right)$$

$$= Pr\left(\frac{\sqrt{n} \cdot \rho_{PZ} \cdot \left(\hat{\beta}_1^{IV} - d + d\right)}{\sqrt{V}} < -z_{1-\frac{\alpha}{2}}\right) + Pr\left(\frac{\sqrt{n} \cdot \rho_{PZ} \cdot \left(\hat{\beta}_1^{IV} - d + d\right)}{\sqrt{V}} > z_{1-\frac{\alpha}{2}}\right)$$

$$= Pr\left(\frac{\sqrt{n} \cdot \rho_{PZ} \cdot \left(\hat{\beta}_1^{IV} - d\right)}{\sqrt{V}} < -z_{1-\frac{\alpha}{2}} - \frac{\sqrt{n} \cdot \rho_{PZ} \cdot d}{\sqrt{V}}\right)$$

$$+Pr\left(\frac{\sqrt{n} \cdot \rho_{PZ} \cdot \left(\hat{\beta}_1^{IV} - d\right)}{\sqrt{V}} > z_{1-\frac{\alpha}{2}} - \frac{\sqrt{n} \cdot \rho_{PZ} \cdot d}{\sqrt{V}}\right)$$

This is an important step because if program impact β_1 is actually d, then

$$\frac{\sqrt{n} \cdot \rho_{PZ} \cdot \left(\hat{\beta}_1^{IV} - d\right)}{\sqrt{V}}$$

has an approximately normal distribution. Thus, if $\beta_1 = d$ we have

$$Pr\left(\frac{\sqrt{n} \cdot \rho_{PZ} \cdot \left(\hat{\beta}_1^{IV} - d\right)}{\sqrt{V}} < -z_{1-\frac{\alpha}{2}} - \frac{\sqrt{n} \cdot \rho_{PZ} \cdot d}{\sqrt{V}}\right)$$

$$+Pr\left(\frac{\sqrt{n} \cdot \rho_{PZ} \cdot \left(\hat{\beta}_1^{IV} - d\right)}{\sqrt{V}} > z_{1-\frac{\alpha}{2}} - \frac{\sqrt{n} \cdot \rho_{PZ} \cdot d}{\sqrt{V}}\right)$$

$$= \Phi\left(-z_{1-\frac{\alpha}{2}} - \frac{\sqrt{n} \cdot \rho_{PZ} \cdot d}{\sqrt{V}}\right) + \left[1 - \Phi\left(z_{1-\frac{\alpha}{2}} - \frac{\sqrt{n} \cdot \rho_{PZ} \cdot d}{\sqrt{V}}\right)\right]$$

Since the start of this process was the derivation of the probability of rejecting the null hypothesis $H_0 : \beta_1 = 0$ if the alternative ($H_a : \beta_1 = d$) is true, this is effectively a power function. In Freeman, et al. they rearrange a bit to

$$Power = 1 - \beta = 1 + \Phi\left(-z_{1-\frac{\alpha}{2}} - \frac{\sqrt{n} \cdot \rho_{PZ} \cdot d}{\sqrt{V}}\right) - \Phi\left(z_{1-\frac{\alpha}{2}} - \frac{\sqrt{n} \cdot \rho_{PZ} \cdot d}{\sqrt{V}}\right)$$

If β_1 is positive, then

$$\Phi\left(-z_{1-\frac{\alpha}{2}} - \frac{\sqrt{n} \cdot \rho_{PZ} \cdot d}{\sqrt{V}}\right)$$

is essentially zero (or at least ignorably small). With the desired power of $1 - \beta$, we can solve for an n of

$$n = \frac{\left(z_{1-\frac{\alpha}{2}} + z_{1-\beta}\right)^2 \cdot V}{d^2 \cdot \rho^2_{PZ}}$$

Given a setup that seems involved in some respect (for example, the two-stage estimation process), it is a refreshing surprise that, compared with the shadows and fog of sample size estimation for the multiple regression, this formula is so simple.

It is probably important to be clear about why it is so simple: There were a limited number of moving parts, despite the complexity of the two-stage least squares model. Some of these need to be addressed.

This model focuses on just one instrument, one outcome and no other exogenous regressors common to the first- and second-stage equations. This is the simplest possible implementation of two-stage least squares, but as such is perhaps the most restrictive version of the linear instrumental variables estimator, in principle and in practice. In classical presentations of linear instrumental variables familiar to most economics graduate students, the discussion of linear instrumental variables typically takes a systems of equations approach, with many potential endogenous variables (or, in the more familiar language we have adopted thus far, outcomes), multiple instruments, and exogenous variables that do not serve as instruments (in the simple two stage least squares example we have focused on, these would be characteristics X_i that appear in both the first- and second-stage equations).

The discussion in this classical setting tends to revolve around whether the system is over-, under-, or just-identified.[122] The issue of identification is typically resolved with linear algebra. By contrast, the single outcome, single instrument discussion of this section was able to avoid linear algebra and focus instead on the comparatively tractable Wald-type estimator. However, if either multiple instruments or exogenous non-instrumental variables X appearing in both the first and second stage had been introduced, the simple framework of the Wald estimator would have been lost, considerably (perhaps fatally) complicating the derivation of a sample size estimator.

Of particular interest for the present purposes is the case of overidentification. In the simple two-stage least squares estimator case (with a single outcome of interest Y and single endogenous variable P), this is the instance where we have more than one first-stage instrument. For instance, rather than one instrument Z_i we might have k instruments $[Z_{1i}, Z_{2i}, ...Z_{ki}]$. There is no easy way to naturally and directly accommodate this in the instrumental variables sample size formula we have developed. Nor would it be easy to extend the formula to accomodate more instruments since the simple distributional approximation for the Wald estimator from which the formula was derived would break down in the case of more than one instrument (indeed, in general, accommodating multiple instruments would probably involve linear algebra and much additional complication).

So how should one decide on a sample size in the presence of the prospect of more than one instrument for program participation? First, it is important to note that this will probably not be a common problem to have. One is typically lucky to have even one plausible instrument for

[122] The simple two-stage least squares setup that motivated the sample size estimation formula developed in this section would be an example of a just-identified system.

program participation. In many cases where one has some *ex-ante* candidate instrument in mind, it will probably be as part of a fallback identification strategy in the setting of an RCT for which there might be less than full compliance.

Nonetheless, there may be circumstances where the study designers believe that more than one instrument for program participation might present itself. An example where this has indeed played out would be a series of papers concerned with the evaluation of a behavior change communications program in Bangladesh (such as Hutchinson, Lance, Guilkey, Shahjahan, & Haque, 2006; Guilkey & Hutchinson, 2011).

For the purpose of *ex-ante* sample size determination, we would recommend not planning on having overidentification. If one thinks about it, sample size estimation that banked on overidentification is effectively sample size estimation that rests on a lot of assumptions: that the program will actually be implemented in a fashion such that each of the candidate instruments will make behavioral sense as an instrument (this is a very big assumption in light of our experiences about how program details evolve even as an impact evaluation of that program is being planned!), that all of the instruments will have useful first stage predictive power, and that all of them will pass overidentification tests (such as the collective Hansen's J-test).

What might be preferable is to base sample size estimation on the assumption that the model will be just-identified in the end (that is, that only one of the candidate instruments will work out in the end). There are a couple of ways to approach this. One could perform sample size estimation for each candidate instrument. In each case one would base estimation on conservative prior assumptions about those instruments. The recommended sample size would then be the maximum among those obtained by considering each instrument individually. One might be able to focus on some subset of the potential instruments about which they possessed particular *ex-ante* confidence.

If one absolutely insists on relying on an assumption of overidentification for sample size estimation, the question then becomes how to accommodate this in the sample size estimation designed for the single instrument case. The most obvious answer is to assume a value for ρ^2_{PZ} based on the regression of P on all of the instruments. This should have the effect of yielding a larger value for ρ^2_{PZ} and hence, other things being equal, a smaller sample size. This is right in line with what one would expect — all other things being equal, the more first-stage explanatory power in the two-stage least squares estimator (in this case exclusively from instruments), the better the performance of the estimator.

A few points need to be emphasized about this model. First, it is motivated by a very simple linear instrumental variables model. As such, it offers little guidance about more complicated instrumental variables estimators. For instance, one might envision using instruments in the setting of bivariate probit estimation, whereby one equation characterizes the program participation decision, and the other the (binary) outcome of interest (estimation of the two equations is typically joint in this case and described as full information estimation; see Lance, Guilkey, et al. [2014] for more details).

Something along the lines of the bivariate probit estimation suggested would represent a (comparatively) complicated nonlinear instrumental variables estimation.[123] The performance of these more complex nonlinear instrumental variables implementations can depend on a variety of factors that might interact in generally complex ways. See Guilkey & Lance (2014) for a Monte Carlo study of some of the factors that influence relative model performance across various sample sizes. In general, we are unaware of neat, closed-form sample size estimation formulas for most (or even, we concede, any) of these more complex extensions of instrumental variables.

That said, even when one plans a more elaborate implementation of instrumental variables, the basic linear model is not a bad framework for thinking about sample size estimation. In general, a conservative approach is again warranted. There is no specific guidance that we can offer (as mentioned, work such as Guilkey and Lance [2014] suggests that model performance in these nonlinear models is complicated and depends on many factors). However, in our own pursuit of such models, we have seen many instances where the stable, precise estimates have required large sample sizes. Clearly there is some limit to this (no one can commit limitless resources to an impact evaluation) but we would recommend either adding sample size to the limit of available resources (which may seem extreme, but one must remember the cost of coming up short on sample size) or, more plausibly, assuming that linear instrumental variables may need to be an eventual fallback position for estimation of impact.

One topic we have not discussed very much is the nature of the outcome Y and endogenous variable P. Program participation could be continuous (as in a dose-response type setup) but more commonly is measured in binary (as in participate/don't participate) terms. The outcome Y could take on virtually any nature, but typically it is either continuous or binary (examples of the latter might include modern contraceptive use, use of antenatal care, school attendance, or pregnancy). A binary endogenous participation variable and binary outcome are exactly the circumstance that could give rise to the aforementioned bivariate probit possibility.

However, linear instrumental variables could also be applied in such a situation. Indeed, linear instrumental variables are frequently recommended in this setting (such as Angrist & Krueger, 2001), often based on concerns regarding misspecification under a nonlinear modelling strategy. Moreover, there is Monte Carlo evidence (for example, Guilkey & Lance, 2014) suggesting that linear instrumental variables performs relatively well in this situation. We therefore would recommend the straightforward linear instrumental sample size estimation approach even in this setting.

Nonetheless, there have been stabs at sample size estimation strategies in nonlinear cases. Burgess (2014) considers logistic regression for the equation of interest (though, again, work such as Angrist and Krueger [2001] would suggest against this practice). Per Burgess, the sample size required to detect an effect size of d per standard deviation increase in P with α significance and $1-\beta$ power is

$$n = \frac{\left(z_{1-\frac{\alpha}{2}} + z_{1-\beta}\right)^2}{d^2 \cdot \rho_{PZ}^2 \cdot P\left(Y=1\right) \cdot P\left(Y=0\right)}$$

[123] Admittedly, given that the nonlinearity alone technically identifies a model along the lines of the bivariate probit that we describe, some would contend that it isn't really instrumental variables. We recuse ourselves from such contentious debates and simply assume that there are some who likely feel otherwise.

It must be emphasized that in the nonlinear case, the effect size d is no longer a classical program impact measure but instead a log odds ratio (or, in other words, an underlying logit parameter).

One important point to remember about the linear instrumental variables based sample size estimator of this section is that it is an approximation. It begins with an approximation, namely the Wald estimator distribution

$$\hat{\beta}_1^{IV} \sim N\left(\beta_1, \frac{var\left(Y|P\right)}{n \cdot var\left(P\right) \cdot \rho_{PZ}^2 \cdot Pr\left(Y=1\right) \cdot Pr\left(Y=0\right)}\right)$$

This is in no sense an exact result. In particular, it likely does not hold even approximately at smaller sample sizes n. What this means in practice is smaller sample size estimates should probably be met with a degree of skepticism. In essence, the formula is producing a sample size estimate in a range for which it was not really designed. This general line of caution is doubly true of extensions of the linear sample size estimator to nonlinear circumstances such as Burgess (2014). These extensions tend to rely as well on approximations for the derivation of the sample that are at least as generous as those used in the linear estimator case.

Before concluding, we should return to the observed variable X_i. We removed this from the model altogether by assuming $\beta_2 = \gamma_1 = 0$. If we had simply assumed the $\beta_2 = 0$ (and left γ_1 in the model), then X would have been ... a potential valid instrument. Whether it would be a good one (in terms of performance) would depend on its power to predict P, which in turn would depend on the magnitude of γ_1. With two instruments (Z_i and X_i) our model would be referred to as "overidentified."

Should one use X as an instrument, in the process estimating an overidentified model? As long as it seems reasonable to expect *ex-ante* that X_i is uncorrelated with μ_i and ϵ_i^Y (in other words, if we believe *ex-ante* that X_i is exogenous), the answer is basically yes. More instruments would mean more first-stage instrumental explanatory power for P. While we cannot naturally characterize this in the sample size estimator introduced above, the logic of the model makes it impossible to believe that this is a bad thing.

Another major benefit to this is that is allowing X to serve as an instrument along with Z allows testing of their validity as instruments. In a linear instrumental variables setting, the only instance where the exogeneity of the instruments (or, put a little bit differently but essentially equivalently, the legitimateness of excluding X and Z from the equation of interest determining Y) can be tested is the overidentified case (see Lance, Guilkey, et al. [2014] for a detailed discussion of such tests). This is another powerful argument for pursuing an overidentified model.

Another interesting case would be where

$$\beta_2 \neq 0 \neq \gamma_1$$

This is the case where X_i is simply another regressor and not necessarily a potential instrument. In this case it is best practice to include X as a regressor in both equations (those determining P and X). The reason is that the performance of the instrumental variables model is likely to improve as overall explanatory power increases.

In conclusion, we would like to point out (what should be) the familiarity of the sample size estimation formula derivation. Although the goal was to develop a sample size estimator for the comparatively complex two-stage least squares instrumental variables estimator,[124] derivation of a sample size estimator was still framed in terms of null and alternative hypotheses, a basic test statistic under the null hypothesis and familiar logic for deriving a power function.

3.5.3 Randomized Controlled Trials and Partial Experiments

One instance in which instrumental variables can come in quite handy is that of a failed RCT. By "failed" we mean that there is less than complete adherence to experimental assignment under the RCT. In this case, some individuals chose their participation status.

To the extent that unobservables such as μ_i have guided the program participation decision (and hence are correlated with P_i), regression of Y on P and X will not yield an unbiased estimation of average program impact β_1. This is the classic endogeneity problem. Put slightly differently, the expected value of μ will vary between participants and non-participants. An RCT with less than complete compliance with experimental assignment will thus yield samples of participants and non-participants that suffer from the very defect (from the standpoint of straightforward unbiased estimation of program impact) that led to the pursuit of an RCT in the first place.

Indeed, there is no remedy for this. An unbiased estimate of β_1 is irretrievably lost. But a consistent estimate of one might be possible via two-stage least sqaures. In particular, as long as there is some degree of compliance with experimental assignment per the RCT, the experimental assignment itself is a kind of valid instrument.

Suppose that A_i is individual i's experimental assignment status. This is defined as

$$A_i = \begin{cases} 1 & \text{if individual } i \text{ was } \textit{assigned} \text{ to participate in the program} \\ 0 & \text{if individual } i \text{ was } \textit{assigned} \text{ not to participate in the program} \end{cases}$$

where assignment is per the RCT design (i.e., the randomization process of the RCT). Estimation then proceeds as described earlier:

1. By regressing P on A and X, estimate by ordinary least squares regression the model:

$$P_i = \kappa_1 + \kappa_2 \cdot A_i + \kappa_3 \cdot X_i + \upsilon_i$$

2. Compute predicted program participation from the fitted model:

$$\hat{P}_i = \hat{\kappa}_1 + \hat{\kappa}_2 \cdot A_i + \hat{\kappa}_3 \cdot X_i$$

3. Regress Y_i on $\hat{P}_i$ and X_i.

[124] Compared with, say, simple comparison of mean outcomes between samples of participants and non-participants.

Notice that A is essentially Z from our earlier introduction of two-stage least squares. (Indeed, we could have written $Z_i = A_i$ and then proceeded exactly as in the earlier enumeration of the steps behind two-stage least squares.) Sometimes an RCT does not exactly involve a clear-cut assignment status A_i but instead the issuance of some kind of random number that somehow informs participation status.

This is very straightforward and offers a significant kind of insurance to those designing RCTs: instrumental variables is a fairly compelling fallback position in the event of incomplete adherence to participation assignment per an RCT. Unfortunately, in the last two decades a growing literature (beginning essentially with Imbens and Angrist [1994]) has identified and explored the implications of an important piece of fine print to this insurance policy.

To understand it, we must expand our basic behavioral model a bit to accommodate heterogeneous program impact, or in other words the possibility that different people experience different levels of program impact. This begins with slightly modified potential outcome equations:

$$Y_i^0 = \beta_0 + \beta_2 \cdot X_i + \beta_3 \cdot \mu_i + \epsilon_i^Y$$

$$Y_i^1 = \beta_0 + \beta_{1i} + \beta_2 \cdot X_i + \beta_3 \cdot \mu_i + \epsilon_i^Y$$

We have an observed (X_i) and unobserved (μ_i) individual-level characteristic determining potential outcomes as well as a purely random, idiosyncratic unobserved component ϵ_i^Y. We assume that the three are independently distributed. Program impact for individual i is

$$Y_i^1 - Y_i^0$$

$$= \beta_0 + \beta_{1i} + \beta_2 \cdot X_i + \beta_3 \cdot \mu_i + \epsilon_i^Y$$
$$- \left(\beta_0 + \beta_2 \cdot X_i + \beta_3 \cdot \mu_i + \epsilon_i^Y\right)$$
$$= \beta_{1i}$$

Notice that program impact is individual-specific. In other words, this model builds in potential for heterogeneous program impact.

A regression specification can be derived in much the same fashion as with behavioral models in preceding chapters. The observed outcome is

$$Y_i = P_i \cdot Y_i^1 + (1 - P_i) \cdot Y_i^0$$

$$= P_i \cdot \left(\beta_0 + \beta_{1i} + \beta_2 \cdot X_i + \beta_3 \cdot \mu_i + \epsilon_i^Y\right)$$
$$+ (1 - P_i) \cdot \left(\beta_0 + \beta_2 \cdot X_i + \beta_3 \cdot \mu_i + \epsilon_i^Y\right)$$
$$= \beta_0 + \beta_{1i} \cdot P_i + \beta_2 \cdot X_i + \beta_3 \cdot \mu_i + \epsilon_i^Y$$

Notice that the unobservable μ_i appears in this equation, and would hence be an element of the regression residual.

The cost of program participation is given by

$$C_i = \gamma_0 + \gamma_1 \cdot X_i + \gamma_{2i} \cdot Z_i + \gamma_4 \cdot \mu_i + \epsilon_i^C$$

We now focus on one instrument, Z_i. For the purposes of this discussion, we generally assume that this instrument is binary. We also assume that $\gamma_{2i} < 0$. This means that an individual becomes more likely to participate as the value of Z_i switches from 0 to 1 because as it does so his/her cost of participation falls by $\gamma_{2i} \cdot Z_i$ (it falls because $\gamma_{2i} < 0$).

Individual i will choose to participate (i.e., his/her value for the program participation indicator P, P_i, equals 1) if

$$Y_i^1 - Y_i^0 - C_i > 0$$

or

$$\beta_{1i} - \gamma_0 - \gamma_1 \cdot X_i - \gamma_{2i} \cdot Z_i - \gamma_4 \cdot \mu_i - \epsilon_i^C > 0$$

where we assume that ϵ_i^C is uncorrelated with both ϵ_i^Y and Z_i. In this framework, the local average treatment effect (LATE) is sometimes presented as involving correlation between β_{1i} and γ_{2i}, the motivation for which is the idea that those most responsive to the instrument (via a large negative value to γ_{2i}) also have a program impact that deviates from the overall population average impact.[125] However, strictly speaking this is not absolutely necessary to show the LATE phenomenon.

In a behavioral setting such as this, it turns out that instrumental variables identifies program impact only for those whose experimental assignment per the RCT determined their participation status. Specifically, we can parse the population and any representative sample from it into three "types":

- **Compliers:** Those whose participated ($P_i = 1$) when they are assigned to do so ($A_i = 1$) and do not participate ($P_i = 0$) when they are not assigned to do so ($A_i = 0$)

- **Never takers:** Those who never participate ($P_i = 0$) regardless of their experimental assignment A_i

- **Always takers:** Those who always participate ($P_i = 1$) regardless of their experimental assignment A_i.

Instrumental variables applied to an RCT with incomplete compliance provides a consistent estimate of average program impact only for compliers, so defined.

The question for present purposes, of course, is what implication this has for sample size estimation. One possibility, of course, would be simply to ignore this possibility for the purpose of sample size estimation. This might be a reasonable course when one has very few priors about the details of LATE in their application.

[125] In their discussion of LATE, Lance, Guilkey, et al. (2014) made such an assumption in part because it showed a more dramatic difference in program impact between compliers and everyone else.

However, there are a few ways in which the possibility of LATE might influence the sample size estimation process. These might include:

1. **First-stage instrument strength:** Other things being equal, the smaller the proportion of the population of interest (and any representative sample from it) that are compliers the smaller will be the the first-stage explanatory power of the instrument. One way of thinking about this is to consider two extremes that one could encounter with a sample of participants and non-participants from an experiment. Consider full compliance (i.e., instrument assignment Z is the only and full determinant of program participation status P). At the other end of the spectrum, the entire population and any representative sample from it could be comprised solely of non-compliers. Presumably this would mean that the population and samples from it are comprised of some mixture of always and never takers (otherwise there would be only participants and no non-participants in the case where everyone was an always taker and no participants in the case where everyone was a never taker). Less than full compliance with experimental assignment has the consequences of reducing the role of experimental assignment status Z in determining program participation P. This translates into lower first stage explanatory power for the instrument.

2. **The "complier" subpopulation:** It is well known (see, for instance, Lance, Guilkey, et al. [2014] and work cited therein) that the broad profile of the complier population is something that can be estimated *ex-post* (i.e., after the "experiment" has been conducted and data on participants and non-participants gathered). However, it may be worth thinking ahead of time about what types of individuals are likely to be compliers. Compliance can be driven by two factors. First, program impact at the individual level plays a role. It seems reasonable that, other things being equal, those with the highest and lowest returns are more likely, other things being equal, to be always and never takers, respectively. Further, it seems likely that, other things being equal, those most responsive to experimental assignment will be most likely to be compliers. That said, there have been many theoretical presentations of LATE, all of which offer different behavioral possibilities in terms of who are likely to be compliers. It is worth thinking about who is most likely to be a complier, and how their average treatment effect might differ from the (best prior guess) about the population average.

Thus LATE can (and should), at least at the margins, inform thinking about the prior guess for the sampling parameters for the sample size estimation formula. Sample size priors should, to the greatest extent possible, be informed by the likely realities of a world of local average treatment effects.

Chapter 4

Conclusion

We end with a brief philosophical note.

A consistent theme in this manual has been its urge for conservatism, by which we have often meant erring on the side of overestimating sample size. However, more generally, conservatism means choosing the sounder path over easy ones. Potentially cost-saving but risky shortcuts or strategies often present themselves throughout the sampling design process. One could use a frame of lesser provenance compared with some more costly but higher-quality alternative. Perhaps there is a sample selection scheme that seems less expensive (for instance, from a fieldwork perspective) than that used by an earlier survey that supplied design-effect estimates and other sample-size estimation priors. Or we might be tempted by the (typically baseless) suggestion that missing a few members of a population won't present detectable bias. We could go on, but the basic idea is clear: sampling design presents many opportunities to cut corners.

Following a more conservative path, either for sample selection or sample-size estimation, tends to lead to larger sample sizes and more cost. When the conservative path yields a sample that turns out to be somewhat too large, the resource waste is on the order of magnitude of "somewhat." But when we do cut corners to drive down costs, we often simply cheat ourselves with sample sizes that prove to be inadequate for meeting the goals of the evaluation, thus wasting *all* of the resources committed.

Moreover, the conservative approach provides us with a better sense of the true costs required to be reasonably confident of meeting our information objectives. We cannot evaluate every program or learn everything we might want to know about the impact of the programs that we do evaluate. In deciding where to concentrate resources, we need to follow some kind of cost-benefit-driven process, and that requires an honest assessment of the costs behind the benefits. Finally, think of your own welfare as samplers: the conservative approach is likely to result in fewer sleepless nights before the survey is completed and the success of the sampling plan can be more comprehensively assessed!

Sometimes we are presented with a situation where an evaluation will proceed no matter what and the available resources do not allow a more conservative approach. In that case, all that the samplers can do is be transparent, by making clear the risks involved in the less-conservative sampling strategy that must be pursued in light of tight resources.

References

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. M. (2014). *Finite population causal standard errors* (No. 20325). Cambridge, MA: National Bureau of Economic Research.

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). Hoboken, N.J.: Wiley-Interscience.

Aiken, A. M., Davey, C., Hargreaves, J. R., & Hayes, R. J. (2015). Re-analysis of health and educational impacts of a school-based deworming programme in western Kenya: A pure replication. *International Journal of Epidemiology*, 44(5), 1572-1580.

Angeles, G., Guilkey, D. K., & Mroz, T. A. (1998). Purposive program placement and the estimation of family planning program effects in Tanzania. *Journal of the American Statistical Association*, 93(443), 884-899.

Angelucci, M., & De Giorgi, G. (2009). Indirect effects of an aid program: How do cash transfers affect ineligibles' consumption? *The American Economic Review*, 486-508.

Angelucci, M. & Di Maro, V. (2015). *Program evaluation and spillover effects* (IZA DP No. 9033). Bonn, Germany: Institute for the Study of Labor.

Angrist, J., & Krueger, A. B. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives*, 15(4), 69-85.

Angrist, J. D., & Pischke, J. S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics*, 119(1), 249-275.

Biemer, P. P., & Christ, S. L. (2008). Weighting survey data. In De Leeuw, E. D., Hox, J. J., & Dillman, D. A. (Eds.), *International handbook of survey methodology* (pp.317-341). New York, NY: Lawrence Erlbaum Associates.

Bierrenbach, A. (2008, August) *Steps in applying probability proportional to size (PPS) and calculating basic probability weights*. Presented at the WHO Training Workshop on TB Prevalence Surveys, Geneva, Switzerland.

Boef, A. G., Dekkers, O. M., Vandenbroucke, J. P., & le Cessie, S. (2014). Sample size importantly limits the usefulness of instrumental variable methods, depending on instrument strength and level of confounding. *Journal of Clinical Epidemiology*, 67(11), 1258-1264.

Bound, J., Jaeger, D. A., & Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association*, 90(430), 443-450.

Burgess, S. (2014). Sample size and power calculations in Mendelian randomization with a single instrumental variable and a binary outcome. *International Journal of Epidemiology*, 43(3), 922-929.

Clemens, M. & Sandefur., J. (2015, July 30). Mapping the worm wars: What the public should take away from the scientific debate about mass deworming [Web log post]. Retrieved from http://www.cgdev.org/blog/mapping-worm-wars-what-public-should-take-away-scientific-debate-about-mass-deworming

Cochran, W.G. (1977). *Sampling techniques* (3rd ed.). New York, NY: Wiley.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (Rev. ed.). New York, NY: Academic Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: L. Erlbaum Associates.

Davey, C., Aiken, A. M., Hayes, R. J., & Hargreaves, J. R. (2015). Re-analysis of health and educational impacts of a school-based deworming programme in western Kenya: A statistical replication of a cluster quasi-randomized stepped-wedge trial. *International Journal of Epidemiology*, 44(5), 1581-1592.

Deaton, A. (2010). Instruments, randomization, and learning about development. *Journal of Economic Literature*, 48(2), 424-455.

Deming, W. (1966). *Some theory of sampling.* New York, NY: Dover.

Dupont, W. D., & Plummer, W. D. (1998). Power and sample size calculations for studies involving linear regression. *Controlled Clinical Trials*, 19(6), 589-601.

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical methods for rates and proportions* (3rd ed.). Hoboken, NJ: Wiley-Interscience.

Freeman, G., Cowling, B. J., & Schooling, C. M. (2013). Power and sample size calculations for Mendelian randomization studies using one genetic instrument. *International Journal of Epidemiology*, 42(4), 1157-1163.

Frison, L., & Pocock, S. J. (1992). Repeated measures in clinical trials: analysis using mean summary statistics and its implications for design. *Statistics in Medicine*, 11(13), 1685-1704.

Gart, J. J. (1985). Approximate tests and interval estimation of the common relative risk in the combination of 2×2 tables. *Biometrika*, 72(3), 673-677.

Gatsonis, C., & Sampson, A. R. (1989). Multiple correlation: Exact power and sample size calculations. *Psychological Bulletin*, 106(3), 516-524.

Gertler, P. J., & Molyneaux, J. W. (1994). How economic development and family planning programs combined to reduce Indonesian fertility. *Demography*, 31(1), 33-63.

Greene, W. H. (2000). *Econometric analysis* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

Greene, W. H. (2004). Fixed effects and bias due to the incidental parameters problem in the Tobit model. *Econometric Reviews*, 23(2), 125-147.

Grau, E., Potter, F., Williams, S., & Diaz-Tena, N. (2006). Nonresponse adjustment using logistic regression: To weight or not to weight? In *Proceedings of the American Statistical Association, Survey Research Methods Section* [CD-ROM]. Alexandria, VA: American Statistical Association.

Guilkey, D. K., & Hutchinson, P. L. (2011). Overcoming methodological challenges in evaluating health communication campaigns: Evidence from rural Bangladesh. *Studies in Family Planning*, 42(2), 93-106.

Guilkey, D. K., & Lance, P. M. (2014). Program impact estimation with binary outcome variables: Monte Carlo results for alternative estimators and empirical examples. In Sickles, R., & Horrace, W. (Eds.), *Festschrift in Honor of Peter Schmidt* (pp. 5-46). New York, NY: Springer.

Hauck Jr, W. W., & Donner, A. (1977). Wald's test as applied to hypotheses in logit analysis.

Journal of the American Statistical Association, 72(360a), 851-853.

Heckman, J. J., Ichimura, H., & Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The Review of Economic Studies*, 64(4), 605-654.

Heckman, J., Ichimura, H., Smith, J., & Todd, P. (1998). Characterizing selection bias using experimental data. *Econometrica*, 66(5), 1017-1098.

Heckman, J. J., & Smith, J. A. (1995). Assessing the case for social experiments. *The Journal of Economic Perspectives*, 9(2), 85-110.

Heeringa, S. G., West, B. T., & Berglund, P. A. (2010). *Applied survey data analysis.* Boca Raton, FL: Chapman & Hall/CRC.

Hemming, K., & Taljaard, M. (2016). Sample size calculations for stepped wedge and cluster randomised trials: A unified approach. *Journal of Clinical Epidemiology*, 69, 137-146.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013) *Applied logistic regression* (3rd ed.). Hoboken, NJ: Wiley.

Hsiao, C. (1986). *Analysis of panel data.* Econometrics Society Monographs. New York, NY: Cambridge University Press.

Hsieh, F. Y. (1989). Sample size tables for logistic regression. *Statistics in medicine*, 8(7), 795-802.

Hsieh, F. Y., Bloch, D. A., & Larsen, M. D. (1998). A simple method of sample size calculation for linear and logistic regression. *Statistics in medicine*, 17(14), 1623-1634.

Hutchinson, P., Lance, P., Guilkey, D. K., Shahjahan, M., & Haque, S. (2006). Measuring the cost-effectiveness of a national health communication program in rural Bangladesh. *Journal of Health Communication*, 11(S2), 91-121.

Iannacchione, V. G. (2003). Sequential weight adjustments for location and cooperation propensity for the 1995 National Survey of Family Growth. *Journal of Official Statistics*, 19(1), 31-43.

Iannacchione, V. G., Milne, J. G., & Folsom, R. E. (1991, August). *Response probability weight adjustments using logistic regression.* Presented at 151st Annual Meeting of the American

Statistical Association, Section on Survey Research Methods, Atlanta, GA.

Imbens, G., & Angrist, J. (1994). Identification and estimation of local average treatment effects. *Econometrica*, 62(2), 467-475.

Jung, S. H., Chow, S. C., & Chi, E. M. (2007). A note on sample size calculation based on propensity analysis in nonrandomized trials. *Journal of Biopharmaceutical Statistics*, 17(1), 35-41.

Kalsbeek, W. D. (1986). Nomad sampling: An analytic study of alternative design strategies. In *Proceedings of the Section on Survey Research Methods, American Statistical Association* (pp. 164-169).

Kalsbeek, W. D., & Cross, A. R. (1982). Problems in sampling nomadic population. In *Proceedings of the Section on Survey Research Methods, American Statistical Assocation* (pp. 398-402).

Kelley, K., & Maxwell, S. E. (2003). Sample size for multiple regression: Obtaining regression coefficients that are accurate, not simply significant. *Psychological Methods*, 8(3), 305-321.

Kish, L. (1965) *Survey sampling.* New York, NY: John Wiley and Sons, Inc.

Kish, L. (1992). Weighting for unequal P_i *Journal of Official Statistics*, 8(2), 183-200.

Korn, E. & Graubard, B. (1999). *Analysis of health surveys.* New York, NY: Wiley.

Lance, P., Guilkey, D., Hattori, A., & Angeles, G. (2014). *How do we know if a program made a difference? A guide to statistical methods for program impact evaluation* (MS-14-87). Chapel Hill, NC: MEASURE Evaluation.

Lance, P., Spencer, J., & Hattori, A. (2014). *GIS and sampling* (MS-14-95). Chapel Hill, NC: MEASURE Evaluation.

Lohr, S. L. (1999). *Sampling: design and analysis.* Pacific Grove, CA: Duxbury Press.

Maddala, G. (1983). *Limited-dependent and qualitative variables in econometrics. Econometric society monographs.* New York, NY: Cambridge University Press.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies. *J Natl Cancer Inst*, 22(4), 719-748.

Maxwell, S. E. (2000). Sample size and multiple regression analysis. *Psychological Methods*, 5(4), 434-458.

McKenzie, D. (2012). Beyond baseline and follow-up: The case for more T in experiments. *Journal of Development Economics*, 99(2), 210-221.

Miguel, E., & Kremer, M. (2004). Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica*, 72(1), 159-217.

Miller, G. (2010). Contraception as development? New evidence from family planning in Colombia. *The Economic Journal*, 120(545), 709-736.

Nam, J. M. (1998). Power and sample size for stratified prospective studies using the score method for testing relative risk. *Biometrics*, 54(1), 331-336.

National Institute of Population Research and Training (NIPORT), Mitra and Associates, and Macro International. (2009). *Bangladesh Demographic and Health Survey 2007.* Dhaka, Bangladesh and Calverton, Maryland, USA: NIPORT, Mitra and Associates, and Macro International.

National Institute of Population Research and Training (NIPORT), MEASURE Evaluation, and icddr,b. (2012). *Bangladesh Maternal Mortality and Health Care Survey 2010.* Dhaka, Bangladesh: NIPORT, MEASURE Evaluation, and icddr,b.

National Institute of Population Research and Training (NIPORT), Mitra and Associates, and ICF International. (2015). *Bangladesh Demographic and Health Survey 2014: Key Indicators.* Dhaka, Bangladesh, and Rockville, Maryland, USA: NIPORT, Mitra and Associates, and ICF International.

Neyman, J., & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, 16(1), 1-32.

Ozler, B. (2015, July 24). Worm wars: A review of the reanalysis of Miguel and Kremers deworming study [Web log post]. Retrieved from http://blogs.worldbank.org/impactevaluations/worm-wars-review-reanalysis-miguel-and-kremer-s-deworming-study

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, 49(12), 1373-1379.

Pitt, M. M., Rosenzweig, M. R., & Gibbons, D. M. (1993). The determinants and consequences of the placement of government programs in Indonesia. *The World Bank Economic Review*, 7(3), 319-348.

Potter, F., Grau, E., Williams, S., Diaz-Tena, N., & Carlson, B. L. (2006). An application of propensity modeling: Comparing unweighted and weighted logistic regression models for nonresponse adjustments. In *Proceedings of the Survey Research Methods Section. American Statistical Association.*

Rao, C. R. (1965). *Linear statistical inference and its applications.* New York, New York: John Wiley & Sons.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41-55.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79(387), 516-524.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1), 33-38.

Rosenzweig, M. R., & Wolpin, K. I. (1986). Evaluating the effects of optimally distributed public programs: Child health and family planning interventions. *The American Economic Review*, 76(3), 470-482.

Rosner, B. (2006) *Fundamentals of biostatistics* (6th ed.). Belmont, CA: Duxbury.

Skinner, C. J., & Rao, J. N. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, 91(433), 349-356.

Staiger, D., & Stock, J. (1997). Instrumental variables regression with weak instruments. *Econometrica*, 65(3), 557-586.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, 25(1), 1-21.

Valliant, R., Dever, J., & Kreuter, F. (2013). *Practical tools for designing and weighting survey samples.* New York, NY: Springer.

Weir, S., Tate, J., Hileman, S., Khan, M., Jackson, E., Johnston, A., & Herman, C. (2005). *Priorities for local AIDS control efforts: A manual for implementing the PLACE method* (MS-05-

13). Chapel Hill, NC: MEASURE Evaluation.

Wooldridge, J. (2006). *Introductory econometrics: A modern approach* (3rd ed.). Mason, OH: Thomson/South-Western.

MEASURE Evaluation
University of North Carolina at Chapel Hill
400 Meadowmont Village Circle, 3rd Floor
Chapel Hill, NC 27517 USA
Phone: +1 919-445-9350
measure@unc.edu www.measureevaluation.org

This publication was produced with the support of the United States Agency for International Development (USAID) under the terms of MEASURE Evaluation cooperative agreement AID-OAA-L-14-00004. MEASURE Evaluation is implemented by the Carolina Population Center, University of North Carolina at Chapel Hill in partnership with ICF International; John Snow, Inc.; Management Sciences for Health; Palladium; and Tulane University. Views expressed are not necessarily those of USAID or the United States government.

USAID
FROM THE AMERICAN PEOPLE

MEASURE
Evaluation