# FIELD EXPERIMENTS

## Design, Analysis, and Interpretation

Alan S. Gerber YALE UNIVERSITY

Donald P. Green COLUMBIA UNIVERSITY

W. W. NORTON & COMPANY NEW YORK • LONDON

W. W. Norton & Company has been independent since its founding in 1923, when William Warder Norton and Mary D. Herter Norton first published lectures delivered at the People's Institute, the adult education division of New York City's Cooper Union. The firm soon expanded its program beyond the Institute, publishing books by celebrated academics from America and abroad. By midcentury, the two major pillars of Norton's publishing program—trade books and college texts—were firmly established. In the 1950s, the Norton family transferred control of the company to its employees, and today—with a staff of four hundred and a comparable number of trade, college, and professional titles published each year—W. W. Norton & Company stands as the largest and oldest publishing house owned wholly by its employees.

Editor: Ann Shin
Associate Editor: Jake Schindel
Project Editor: Jack Borrebach
Marketing Manager, political science: Sasha Levitt
Production Manager: Eric Pier-Hocking
Text Design: Joan Greenfield / Gooddesign Resource
Design Director: Hope Miller Goodell
Composition by Jouve International—Brattleboro, VT
Manufacturing by the Maple Press—York, PA


Printed in the United States of America.
First edition.

**Library of Congress Cataloging-in-Publication Data**

Gerber, Alan S.
Field experiments : design, analysis, and interpretation / Alan S. Gerber, Donald P. Green. — 1st ed.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-393-97995-4 (pbk.)
1. Political science—Research—Methodology. 2. Social science—Research—Methodology.
3. Political science—Study and teaching (Higher) 4. Social science—Study and teaching (Higher)
I. Green, Donald P., 1961-II. Title.
JA86.G36 2012
001.4'34—dc23

2011052337

W. W. Norton & Company, Inc., 500 Fifth Avenue, New York, NY 10110-0017
wwnorton.com
W. W. Norton & Company Ltd., Castle House, 75/76 Wells Street, London W1T 3QT

THIS BOOK IS DEDICATED TO OUR PARENTS, WHO HELPED INSTILL IN US A LOVE OF SCIENCE.

# CONTENTS

# CHAPTER 1

# Introduction

Daily life continually presents us with questions of cause and effect. Will eating more vegetables make me healthier? If I drive a bit faster than the law allows, will the police pull me over for a speeding ticket? Will dragging my reluctant children to museums make them one day more interested in art and history? Even actions as banal as scheduling a dental exam or choosing an efficient path to work draw on cause-and-effect reasoning.

Organizations, too, grapple with causal puzzles. Charities try to figure out which fundraising appeals work best. Marketing agencies look for ways to boost sales. Churches strive to attract congregants on Sundays. Political parties maneuver to win elections. Interest groups attempt to influence legislation. Whether their aim is to boost donations, sales, attendance, or political influence, organizations make decisions based (at least in part) on their understanding of cause and effect. In some cases, the survival of an organization depends on the skill with which it addresses the causal questions that it confronts.

Of special interest to academic researchers are the causal questions that confront governments and policy makers. What are the economic and social effects of raising the minimum wage? Would allowing parents to pay for private school using publicly funded vouchers make the educational system more effective and cost-efficient? Would legal limits on how much candidates can spend when running for office affect the competitiveness of elections? In the interest of preventing bloodshed, should international peacekeeping troops be deployed with or without heavy weapons? Would mandating harsher punishments for violent offenders deter crime? A list of policy-relevant causal questions would itself fill a book.

An even larger tome would be needed to catalog the many theoretical questions that are inspired by causal claims. For example, when asked to contribute to a collective cause, such as cutting down on carbon emissions in order to prevent global climate change, to what extent are people responsive to appeals based on social norms or ideology? Prominent scholars have argued that collective action will founder

unless individuals are given some sort of reward for their participation; according to this argument, simply telling people that they ought to contribute to a collective cause will not work.[1] If this underlying causal claim is true, the consequences for policymaking are profound: tax credits may work, but declaring a national Climate Change Awareness Day will not.

Whether because of their practical, policy, or theoretical significance—or simply because they transport us to a different time and place—causal claims spark the imagination. How does the pilgrimage to Mecca affect the religious, social, and political attitudes of Muslims?[2] Do high school dropout rates in low-income areas improve when children are given monetary rewards for academic performance?[3] Are Mexican police more likely to demand bribes from upper- or lower-class drivers who are pulled aside for traffic infractions?[4] Does your race affect whether employers call you for a job interview?[5] In the context of a civil war, do civilians become more supportive of the government when local economic conditions improve?[6] Does artillery bombardment directed against villages suspected of harboring insurgent guerrillas increase or decrease the likelihood of subsequent insurgent attacks from those villages?[7]

In short, the world is brimming over with causal questions. How might one go about answering them in a convincing manner? What methods for answering causal questions should be viewed with skepticism?

## 1.1 Drawing Inferences from Intuitions, Anecdotes, and Correlations

One common way of addressing causal questions is to draw on intuition and anecdotes. In the aforementioned case of artillery directed at insurgent villages, a scholar might reason that firing on these villages could galvanize support for the rebels, leading to more insurgent attacks in the future. Bombardment might also prompt the rebels to demonstrate to villagers their determination to fight on by escalating their insurgent activities. In support of this hypothesis, one might point out that the anti-Nazi insurgency in Soviet Russia in 1941 became more determined after occupation forces stepped up their military suppression. One problem with building causal arguments

1 Olson 1965.
2 Clingingsmith, Khwaja, and Kremer 2009.
3 Angrist and Lavy 2009; see also Fryer 2010.
4 Fried, Lagunes, and Venkataramani 2010.
5 Bertrand and Mullainathan 2004.
6 Beath, Christia, and Enikolopov 2011.
7 Lyall 2009.



around intuitions and anecdotes, however, is that such arguments can often be adduced for both sides of a causal claim. In the case of firing on insurgents, another researcher could argue that insurgents depend on the goodwill of villagers; once a village is fired upon, villagers have a greater incentive to expel the rebels in order to prevent future attacks. Supplies dry up, and informants disclose rebel hideouts to government forces. This researcher could defend the argument by describing the government suppression of the Sanusi uprising in Libya, which seemed to deal a lasting blow to these rebels' ability to carry out insurgent attacks.[8] Debates based on intuition and anecdotes frequently result in stalemate.

A critique of anecdote and intuition can be taken a step further. The method is susceptible to error even when intuition and anecdotes seem to favor just one side of an argument. The history of medicine, which is instructive because it tends to provide clearer answers to causal questions than research in social science, is replete with examples of well-reasoned hypotheses that later proved to be false when tested experimentally. Consider the case of aortic arrhythmia (irregular heartbeat), which is often associated with heart attacks. A well-regarded theory held that arrhythmia was a precursor to heart attack. Several drugs were developed to suppress arrhythmia, and early clinical reports seemed to suggest the benefits of restoring a regular heartbeat. The Cardiac Arrhythmia Suppression Trial, a large randomized experiment, was launched in the hope of finding which of three suppression drugs worked best, only to discover that two of the three drugs produced a significant *increase* in death and heart attacks, while the third had negative but seemingly less fatal consequences.[9] The broader point is that well-regarded theories are fallible. This concern is particularly acute in the social sciences, where intuitions are rarely uncontroversial, and controversial intuitions are rarely backed up by conclusive evidence.

Another common research strategy is to assemble statistical evidence showing that an outcome becomes more likely when a certain cause is present. Researchers sometimes go to great lengths to assemble large datasets that allow them to track the correlation between putative causes and effects. These data might be used to learn about the following statistical relationship: to what extent do villages that come under attack by government forces tend to have more or less subsequent insurgent activity? Sometimes these analyses turn up robust correlations between interventions and outcomes. The problem is that correlations can be a misleading guide to causation. Suppose, for example, that the correlation between government bombardment and subsequent insurgent activity were found to be strongly positive: the more shelling, the more subsequent insurgent activity. If interpreted causally, this correlation would indicate that shelling prompted insurgents to step up their attacks. Other

8 See Lyall 2009 for a discussion of these debates and historical episodes.
9 Cardiac Arrhythmia Suppression Trial II Investigators 1992.

interpretations, however, are possible. It could be that government forces received intelligence about an escalation of insurgent activity in certain villages and directed their artillery there. Shelling, in other words, could be a marker for an uptick in insurgent activity. Under this scenario, we would observe a positive correlation between shelling and subsequent insurgent attacks even if shelling per se had no effect.

The basic problem with using correlations as a guide to causality is that correlations may arise for reasons that have nothing to do with the causal process under investigation. Do SAT preparation courses improve SAT scores? Suppose there were a strong positive correlation here: people who took a prep class on average got higher SAT scores than those who did not take the prep class. Does this correlation reflect the course-induced improvement in scores, or rather the fact that students with the money and motivation to take a prep course tend to score higher than their less affluent or less motivated counterparts? If the latter were true, we might see a strong association even if the prep course had no effect on scores. A common error is to reason that where there's smoke, there's fire: correlations at least hint at the existence of a causal relationship, right? Not necessarily. Basketball players tend to be taller than other people, but you cannot grow taller by joining the basketball team.

The distinction between correlation and causation seems so fundamental that one might wonder why social scientists rely on correlations when making causal arguments. The answer is that the dominant methodological practice is to transform raw correlations into more refined correlations. After noticing a correlation that might have a causal interpretation, researchers attempt to make this causal interpretation more convincing by limiting the comparison to observations that have similar background attributes. For example, a researcher seeking to isolate the effects of the SAT preparatory course might restrict attention to people with the same gender, age, race, grade point average, and socioeconomic status. The problem is that this method remains vulnerable to *unobserved* factors that predict SAT scores and are correlated with taking a prep course. By restricting attention to people with the same socio-demographic characteristics, a researcher makes the people who took the course comparable to those who did not in terms of *observed* attributes, but these groups may nevertheless differ in ways that are unobserved. In some cases, a researcher may fail to consider some of the factors that contribute to SAT scores. In other cases, a researcher may think of relevant factors but fail to measure them adequately. For example, people who take the prep course may, on average, be more motivated to do well on the test. If we fail to measure motivation (or fail to measure it accurately), it will be one of the unmeasured attributes that might cause us to draw mistaken inferences. These unmeasured attributes are sometimes called *confounders* or *lurking variables* or *unobserved heterogeneity*. When interpreting correlations, researchers must always be alert to the distorting influence of unmeasured attributes. The fact that someone chooses to take the prep course may reveal something about how they are likely to perform on the test. Even if the course truly has no



effect, people with the same age, gender, and affluence may seem to do better when they take the course.

Whether the problem of unobserved confounders is severe or innocuous will depend on the causal question at hand and the manner in which background attributes are measured. Consider the so-called "broken windows" theory, which suggests that crime increases when blighted areas appear to be abandoned by property owners and unsupervised by police.[10] The causal question is whether one could reduce crime in such areas by picking up trash, removing graffiti, and repairing broken windows. A weak study might compare crime rates on streets with varying levels of property disrepair. A more convincing study might compare crime rates on streets that currently experience different levels of blight but in the past had similar rates of disrepair and crime. But even the latter study may still be unconvincing because unmeasured factors, such as the closing of a large local business, may have caused some streets to deteriorate physically and coincided with an upsurge in crime.[11]

Determined to conquer the problem of unobserved confounders, one could set out to measure each and every one of the unmeasured factors. The intrepid researcher who embarks on this daunting task confronts a fundamental problem: no one can be sure what the set of unmeasured factors comprises. The list of all potential confounders is essentially a bottomless pit, and the search has no well-defined stopping rule. In the social sciences, research literatures routinely become mired in disputes about unobserved confounders and what to do about them.

## 1.2 Experiments as a Solution to the Problem of Unobserved Confounders

The challenge for those who seek to answer causal questions in a convincing fashion is to come up with a research strategy that does not require them to identify, let alone measure, all potential confounders. Gradually, over the course of centuries, researchers developed procedures designed to sever the statistical relationship between the treatment and all variables that predict outcomes. The earliest experiments, such as Lind's study of scurvy in the 1750s and Watson's study of smallpox in the 1760s, introduced the method of systematically tracking the effects of a researcher-induced intervention by comparing outcomes in the treatment group to outcomes in one or more control groups.[12] One important limitation of these early studies is that they assumed that their subjects were identical in terms of their medical trajectories. What if this assumption

10 Wilson and Kelling 1982.

11 See Keizer, Lindenberg, and Steg 2008, but note that this study does not employ random assignment. For a randomized field experiment see Mazerolle, Price, and Roehl 2000.

12 Hughes 1975; Boylston 2008.

were false, and treatments tended to be administered to patients with the best chances of recovery? Concerned that the apparent effects of their intervention might be attributable to extraneous factors, researchers placed increasing emphasis on the procedure by which treatments were assigned to subjects. Many pathbreaking studies of the nineteenth century assigned subjects alternately to treatment and control in an effort to make the experimental groups comparable. In 1809, a Scottish medical student described research conducted in Portugal in which army surgeons treated 366 sick soldiers alternately with bloodletting and other palliatives.[13] In the 1880s, Louis Pasteur tested his anthrax vaccine on animals by alternately exposing treatment and control groups to the bacteria. In 1898, Johannes Fibiger assigned an experimental treatment to diphtheria patients admitted to a hospital in Copenhagen on alternate days.[14] Alternating designs were common in early agricultural studies and investigations of clairvoyance, although researchers gradually came to recognize potential pitfalls of alternation.[15] One problem with alternating designs is that they cannot definitively rule out confounding factors, such as sicker diphtheria patients coming to the hospital on certain days of the week. The first to recognize the full significance of this point was the agricultural statistician R. A. Fisher, who in the mid-1920s argued vigorously for the advantages of assigning observations at random to treatment and control conditions.[16]

This insight represents a watershed moment in the history of science. Recognizing that no planned design, no matter how elaborate, could fend off every possible systematic difference between the treatment and control groups, Fisher laid out a general procedure for eliminating systematic differences between treatment and control groups: random assignment. When we speak of experiments in this volume, we refer to studies in which some kind of random procedure, such as a coin flip, determines whether a subject receives a treatment.

One remarkable aspect of the history of randomized experimentation is that the idea of random assignment occurred to several ingenious people centuries before it was introduced into modern scientific practice. For example, the notion that one could use random assignment to form comparable experimental groups seems to have been apparent to the Flemish physician Jan Baptist Van Helmont, whose 1648 manuscript "Origin of Medicine" challenged the proponents of bloodletting to perform the following randomized experiment:

> Let us take out of the hospitals . . . 200 or 500 poor people, that have fevers, pleurisies. Let us divide them into halves, let us cast lots, that one halfe of them may fall to my share, and the other to yours; I will cure them without bloodletting and sensible evacuation; but you do, as ye know . . . We shall see how many funerals both of us shall have.[17]

13 Chalmers 2001.

14 Hróbjartsson, Gøtzsche, and Gluud 1998.

15 Merrill 2010. For further reading on the history of experimentation, see Cochran 1976; Forsetlund, Chalmers, and Bjørndal 2007; Hacking 1990; and Salsburg 2001. See Greenberg and Shroder 2004 on social experiments and Green and Gerber 2003 on the history of experiments in political science.

16 Box 1980, p. 3.



Unfortunately for those whose physicians prescribed bloodletting in the centuries following Van Helmont, he never conducted his proposed experiment. One can find similar references to hypothetical experiments dating back to medieval times, but no indication that any were actually put into practice. Until the advent of modern statistical theory in the early twentieth century, the properties of random assignment were not fully appreciated, nor were they discussed in a systematic manner that would have allowed one generation to recommend the idea to the next.

Even after Fisher's ideas became widely known in the wake of his 1935 book *The Design of Experiments*, randomized designs met resistance from medical researchers until the 1950s, and randomized experiments did not catch on in the social sciences until the 1960s.[18] In the class of brilliant twentieth-century discoveries, the idea of randomization contrasts sharply with the idea of relativity, which lay completely hidden until uncovered by genius. Randomization was more akin to crude oil, something that periodically bubbled to the surface but remained untapped for centuries until its extraordinary practical value came to be appreciated.

## 1.3 Experiments as Fair Tests

In the contentious world of causal claims, randomized experimentation represents an evenhanded method for assessing what works. The procedure of assigning treatments at random ensures that there is no systematic tendency for either the treatment or control group to have an advantage. If subjects were assigned to treatment and control groups and no treatment were actually administered, there would be no reason to expect that one group would outperform the other. In other words, random

17 Chalmers 2001, p. 1157.

18 The advent of randomized experimentation in social and medical research took roughly a quarter century. Shortly after laying the statistical foundations for random assignment and the analysis of experimental data, Fisher collaborated on the first randomized agricultural experiment (Eden and Fisher 1927). Within a few years, Amberson, McMahon, and Pinner (1931) performed what appears to be the first randomized medical experiment, in which tuberculosis patients were assigned to clinical trials based on a coin flip. The large-scale studies of tuberculosis conducted during the 1940s brought randomized clinical trials to the forefront of medicine. Shortly afterward, the primacy of this methodology in medicine was cemented by a series of essays by Hill (1951, 1952) and subsequent acclaim of the polio vaccine trials of the 1950s (Tanur 1989). Randomized clinical trials gradually came to be heralded as the gold standard by which medical claims were to be judged. By 1952, books such as Kempthorne's *Design and Analysis of Experiments* (pp. 125–126) declared that "only when the treatments in the experiment are applied by the experimenter using the full randomization procedure is the chain of inductive inference sound."

assignment implies that the observed *and unobserved* factors that affect outcomes are equally likely to be present in the treatment and control groups. Any given experiment may overestimate or underestimate the effect of the treatment, but if the experiment were conducted repeatedly under similar conditions, the average experimental result would accurately reflect the true treatment effect. In Chapter 2, we will spell out this feature of randomized experiments in greater detail when we discuss the concept of unbiased estimation.

Experiments are fair in another sense: they involve transparent, reproducible procedures. The steps used to conduct a randomized experiment may be carried out by any research group. A random procedure such as a coin flip may be used to allocate observations to treatment or control, and observers can monitor the random assignment process to make sure that it is followed faithfully. Because the allocation process precedes the measurement of outcomes, it is also possible to spell out beforehand the way in which the data will be analyzed. By automating the process of data analysis, one limits the role of discretion that could compromise the fairness of a test.

Random allocation is the dividing line that separates experimental from nonexperimental research in the social sciences. When working with nonexperimental data, one cannot be sure whether the treatment and control groups are comparable because no one knows precisely why some subjects and not others came to receive the treatment. A researcher may be prepared to assume that the two groups are comparable, but assumptions that seem plausible to one researcher may strike another as far-fetched.

This is not to say that experiments are free from problems. Indeed, this book would be rather brief were it not for the many complications that may arise in the course of conducting, analyzing, and interpreting experiments. Entire chapters are devoted to problems of noncompliance (subjects who receive a treatment other than the one to which they were randomly assigned), attrition (observations for which outcome measurements are unavailable), and interference between units (observations influenced by the experimental conditions to which other observations are assigned). The threat of bias remains a constant concern even when conducting experiments, which is why it is so important to design and analyze them with an eye toward maintaining symmetry between treatment and control groups and, more generally, to embed the experimental enterprise in institutions that facilitate proper reporting and accumulation of experimental results.

## 1.4 Field Experiments

Experiments are used for a wide array of different purposes. Sometimes the aim of an experiment is to assess a theoretical claim by testing an implied causal relationship. Game theorists, for example, use laboratory experiments to show how the introduction



**BOX 1.1**

### Experiments in the Natural Sciences

Readers with a background in the natural sciences may find it surprising that random assignment is an integral part of the definition of a social science experiment. Why is random assignment often unnecessary in experiments in, for example, physics? Part of the answer is that the "subjects" in these experiments—e.g., electrons—are more or less interchangeable, and so the method used to assign subjects to treatment is inconsequential. Another part of the answer is that lab conditions neutralize all forces other than the treatment.

In the life sciences, subjects are often different from one another, and eliminating unmeasured disturbances can be difficult even under carefully controlled conditions. An instructive example may be found in a study by Crabbe, Wahlsten, and Dudek (1999), who performed a series of experiments on mouse behavior in three different science labs. As Lehrer (2010) explains:

> Before [Crabbe] conducted the experiments, he tried to standardize every variable he could think of. The same strains of mice were used in each lab, shipped on the same day from the same supplier. The animals were raised in the same kind of enclosure, with the same brand of sawdust bedding. They had been exposed to the same amount of incandescent light, were living with the same number of littermates, and were fed the exact same type of chow pellets. When the mice were handled, it was with the same kind of surgical glove, and when they were tested it was on the same equipment, at the same time in the morning.

Nevertheless, experimental interventions produced markedly different results across mice and research sites.

of uncertainty or the opportunity to exchange information prior to negotiating affects the bargains that participants strike with one another.[19] Such experiments are often couched in very abstract terms, with rules that stylize the features of an auction, legislative session, or international dispute. The participants are typically ordinary people (often members of the university community), not traders, legislators, or diplomats, and the laboratory environment makes them keenly aware that they are participating in a research study.

At the other end of the spectrum are experiments that strive to be as realistic and unobtrusive as possible in an effort to test more context-specific hypotheses.

19 See Davis and Holt 1993; Kagel and Roth 1995; Guala 2005.

Quite often this type of research is inspired by a mixture of theoretical and practical concerns. For example, to what extent and under what conditions does preschool improve subsequent educational outcomes? Experiments that address this question shed light on theories about childhood development while at the same time informing policy debates about whether and how to allocate resources to early childhood education in specific communities.

The push for realism and unobtrusiveness stems from the concern that unless one conducts experiments in a naturalistic setting and manner, some aspect of the experimental design may generate results that are idiosyncratic or misleading. If subjects know that they are being studied or if they sense that the treatment they received is supposed to elicit a certain kind of response, they may express the opinions or report the behavior they believe the experimenter wants to hear. A treatment may seem effective until a more unobtrusive experiment proves otherwise.[20] Conducting research in naturalistic settings may be viewed as a hedge against unforeseen threats to inference that arise when drawing generalizations from results obtained in laboratory settings. Just as experiments are designed to test causal claims with minimal reliance on assumptions, experiments conducted in real-world settings are designed to make generalizations less dependent on assumptions.

Randomized studies that are conducted in real-world settings are often called *field experiments*, a term that calls to mind early agricultural experiments that were literally conducted in fields. The problem with the term is that the word *field* refers to the setting, but the setting is just one aspect of an experiment. One should invoke not one but several criteria: whether the treatment used in the study resembles the intervention of interest in the world, whether the participants resemble the actors who ordinarily encounter these interventions, whether the context within which subjects

20 Whether this concern is justified is an empirical question, and the answer may well depend on the setting, context, and subjects. Unfortunately, the research literature on this topic remains underdeveloped. Few studies have attempted to estimate treatment effects in both lab and field contexts. Gneezy, Haruvy, and Yafe (2004), for example, use field and lab studies to test the hypothesis that the quantity of food consumed depends on whether each diner pays for his or her own food or whether they all split the bill. When this experiment is conducted in an actual cafeteria, splitting the bill leads to significantly more food consumption; when the equivalent game is played in abstract form (with monetary payoffs) in a nearby lab, the average effect is weak and not statistically distinguishable from zero. Jerit, Barabus, and Clifford (2012) compare the effects of exposure to a local newspaper on political knowledge and opinions. In the field, free Sunday newspapers were randomly distributed to households over the course of one month; in the lab, subjects from the same population were invited to a university setting, where they were presented with the four most prominent political news stories airing during the same month. For the 17 outcome measures, estimated treatment effects in the lab and field are found to be weakly correlated (Table 2). See also Rondeau and List (2008), who compare the effectiveness of different fundraising appeals on behalf of the Sierra Club directed at 3,000 past donors, as measured by actual donations. The fundraising appeals, which involve various combinations of matching funds, thresholds, and money-back guarantees, are then presented in abstract form in a lab setting with monetary payoffs. The correspondence between lab and field results was relatively weak, with average contributions in the lab predicting about 5% of the variance in average contributions in the field across the four conditions.



receive the treatment resembles the context of interest, and whether the outcome measures resemble the actual outcomes of theoretical or practical interest.

For example, suppose one were interested in the extent to which financial contributions to incumbent legislators' reelection campaigns buy donors access to the legislators, a topic of great interest to those concerned that the access accorded to wealthy donors undermines democratic representation. The hypothesis is that the more a donor contributes, the more likely the legislator is to grant a meeting to discuss the donor's policy prescriptions. One possible design is to recruit students to play the part of legislative schedulers and present them with a list of requests for meetings from an assortment of constituents and donors in order to test whether people described as potential donors receive priority. Another design involves the same exercise, but this time the subjects are actual legislative schedulers.[21] The latter design would seem to provide more convincing evidence about the relationship between donations and access in actual legislative settings, but the degree of experimental realism remains ambiguous. The treatments in this case are realistic in the sense that they resemble what an actual scheduler might confront, but the subjects are aware that they are participating in a simulation exercise. Under scrutiny by researchers, legislative schedulers might try to appear indifferent to fundraising considerations; in an actual legislative setting where principals provide feedback to schedulers, donors might receive special consideration. More realistic, then, would be an experiment in which one or more donors contribute randomly assigned sums of money to various legislators and request meetings to discuss a policy or administrative concern. In this design, the subjects are actual schedulers, the treatment is a campaign donation, the treatment and request for a meeting are authentic, and the outcome is whether a real request is granted in a timely fashion.

Because the degree of "fieldness" may be gauged along four different dimensions (authenticity of treatments, participants, contexts, and outcome measures), a proper classification scheme would involve at least sixteen categories, a taxonomy that far exceeds anyone's interest or patience. Suffice it to say that field experiments take many forms. Some experiments seem naturalistic on all dimensions. Sherman et al. worked with the Kansas City police department in order to test the effectiveness of police raids on locations where drug dealing was suspected.[22] The treatments were raids by teams of uniformed police directed at 104 randomly chosen sites among the 207 locations for which warrants had been issued. Outcomes were crime rates in nearby areas. Karlan and List collaborated with a charity in order to test the effectiveness of alternative fundraising appeals.[23] The treatments were fundraising letters; the experiment was unobtrusive in the sense that recipients of the fundraising appeals were

21 See Chin, Bond, and Geva 2000.

22 Sherman et al. 1995.

23 Karlan and List 2007.

unaware that an experiment was being conducted; and the outcomes were financial donations. Bergan teamed up with a grassroots lobbying organization in order to test whether constituents' e-mail to state representatives influences roll call voting.[24] The lobbying organization allowed Bergan to extract a random control group from its list of targeted legislators; otherwise, its lobbying campaign was conducted in the usual way, and outcomes were assessed based on the legislators' floor votes.

Many field experiments are less naturalistic, and generalizations drawn from them are more dependent on assumptions. Sometimes the interventions deployed in the field are designed by researchers rather than practitioners. Eldersveld, for example, fashioned his own get-out-the-vote campaigns in order to test whether mobilization activities cause registered voters to cast ballots.[25] Much may be learned when researchers craft their own treatments—indeed, the development of theoretically inspired interventions is an important way in which researchers may contribute to theoretical and policy debates. However, if the aim of an experiment is to gauge the effectiveness of typical candidate- or party-led voter mobilization campaigns, researcher-led campaigns may be unrepresentative in terms of the messages used or the manner in which they are communicated. Suppose that the researcher's intervention were to prove ineffective. This finding alone would not establish that a typical campaign's interventions are ineffective, although this interpretation could be bolstered by a series of follow-up experiments that test different types of campaign communication.[26] Sometimes treatments are administered and outcomes are measured in a way that notifies participants that they are being studied, as in Paluck's experimental investigation of intergroup prejudice in Rwanda.[27] Her study enlisted groups of Rwandan villagers to listen to recordings of radio programs on a monthly basis for a period of one year, at which point outcomes were measured using surveys and role-playing exercises. Finally, experimental studies with relatively little field content are those in which actual interventions are delivered in artificial settings to subjects who are aware that they are part of a study. Examples of this type of research may be found in the domain of commercial advertising, where subjects are shown different types of ads either in the context of an Internet survey or in a lab located in a shopping center.[28]

Whether a given study is regarded as a field experiment is partly a matter of perspective. Ordinarily, experiments that take place on college campuses are considered lab studies, but some experiments on cheating involve realistic opportunities for students to copy answers or misreport their own performance on self-graded tests.[29] An experimental study that examines the deterrent effect of exam proctoring would amount to a field experiment if one's aim were to understand the conditions under which students cheat in school. This example serves as a reminder that what constitutes a field experiment depends on how "the field" is defined.

24 Bergan 2009.

25 Eldersveld 1956.

26 For example, in an effort to test whether voter mobilization phone calls conducted by call centers are typically ineffective, Panagopoulos (2009) compares partisan and nonpartisan scripts, Nickerson (2007) assesses whether effectiveness varies depending on the quality of the calling center, and other scholars have conducted studies in various electoral environments. See Green and Gerber 2008 for a review of this literature.

27 Paluck 2009.

28 See, for example, Clinton and Lapinski 2004; Kohn, Smart, and Ogborne 1984.



## 1.5 Advantages and Disadvantages of Experimenting in Real-World Settings

Many field experiments take the form of "program evaluations" designed to gauge the extent to which resources are deployed effectively. For example, in order to test whether a political candidate's TV advertising campaign increases her popularity, a field experiment might randomize the geographic areas in which the ads are deployed and measure differences in voter support between treatment and control regions. From the standpoint of program evaluation, this type of experiment is arguably superior to a laboratory study in which voters are randomly shown the candidate's ads and later asked their views about the candidate. The field experiment tests the effects of deploying the ads and allows for the possibility that some voters in targeted areas will miss the ad, watch it inattentively, or forget its message amid life's other distractions. Interpretation of the lab experiment's results is complicated by the fact that subjects in lab settings may respond differently to the ads than the average voter outside the lab. In this application, preliminary lab research might be useful insofar as it suggests which messages are most likely to work in field settings, but only a field experiment allows the researcher to reliably gauge the extent to which an actual ad campaign changed votes and to express this outcome in relation to the resources spent on the campaign.

As we move from program evaluation to tests of theoretical propositions, the relative merits of field and lab settings become less clear-cut. A practical advantage of delivering treatments under controlled laboratory conditions is that one can more easily administer multiple variations of a treatment to test fine-grained theoretical propositions. Field interventions are often more cumbersome: in the case of political advertisements, it may be logistically challenging or politically risky to air multiple advertisements in different media markets. On the other hand, field experiments are sometimes able to achieve a high level of theoretical nuance when a wide array of treatments can be distributed across a large pool of subjects. Field experiments that deploy multiple versions of a treatment are common, for example, in research

29 Canning 1956; Nowell and Laufer 1997.

on discrimination, where researchers vary ethnicity, social class, and a host of other characteristics to better understand the conditions under which discrimination occurs.[30]

Even when limited to a single, relatively blunt intervention, a researcher may still have reason to conduct experiments in the field. Advertising research in field settings is often unobtrusive in the sense that subjects are not viewing the ad at the behest of a researcher, and outcomes are measured in a way that does not alert subjects to the fact that they are being studied.[31] Whereas outcomes in lab settings are often attitudes and behaviors that can be measured in the space of one sitting,[32] field studies tend to monitor behaviors over extended periods of time. The importance of ongoing outcome measurement is illustrated by experiments that find strong instantaneous effects of political advertising that decay rapidly over time.[33]

Perhaps the biggest disadvantage of conducting experiments in the field is that they are often challenging to implement. In contrast to the lab, where researchers can make unilateral decisions about what treatments to deploy, field experiments are often the product of coordination between researchers and those who actually carry out the interventions or furnish data on subjects' outcomes. Orr[34] and Gueron[35] offer helpful descriptions of how these partnerships are formed and nurtured over the course of a collaborative research project. Both authors stress the importance of building consensus about the use of random assignment. Research partners and funders sometimes balk at the idea of randomly allocating treatments, preferring instead to treat everyone or a hand-picked selection of subjects. The researcher must be prepared to formulate a palatable experimental design and to argue convincingly that the proposed use of random assignment is both feasible and ethical. The authors also stress that successful implementation of the agreed-upon experimental design—the allocation of subjects, the administration of treatments, and the measurement of outcomes—requires planning, pilot testing, and constant supervision.

Managing research collaboration with schools, police departments, retail firms, or political campaigns sounds difficult and often is. Nevertheless, field experimentation is a rapidly growing form of social science research, encompassing hundreds of

30 See Doleac and Stein 2010 for a study of racial discrimination by bidders on Internet auctions or Pager, Western, and Bonikowski 2009 for a study of labor market discrimination. We discuss discrimination experiments in Chapters 9 and 12.

31 In cases where surveys are used to assess outcomes, measurement may be unobtrusive in the more limited but nevertheless important sense that subjects are unaware that the survey aims to gauge the effects of the intervention.

32 Orchestrating return visits to the lab often presents logistical challenges, and failure to attract all subjects back to the lab may introduce bias (see Chapter 7).

33 See, for example, Gerber, Gimpel, Green, and Shaw 2011. See also the discussion of outcome measurement in Chapter 12.

34 Orr 1999, Chapter 5.

35 Gueron 2002.



studies on topics like education, crime, employment, savings, discrimination, charitable giving, conservation, and political participation.[36] The set of noteworthy and influential studies includes experiments of every possible description: small-scale interventions designed and implemented by researchers; collaborations between researchers and firms, schools, police agencies, or political campaigns; and massive government-funded studies of income taxes, health insurance, schooling, and public housing.[37]

Time and again, researchers overcome practical hurdles, and the boundaries of what is possible seem to be continually expanding. Consider, for example, research on how to promote government accountability. Until the 1990s, research in this domain was almost exclusively nonexperimental, but a series of pathbreaking studies have shown that one can use experiments to investigate the effects of government audits and community forums on accounting irregularities among public works programs,[38] the effects of grassroots monitoring efforts on the performance of legislators,[39] and the effects of information about constituents' preferences on legislators' roll call votes.[40] Field experiments are sometimes faulted for their inability to address big questions, such as the effects of culture, wars, or constitutions, but researchers have grown increasingly adept at designing experiments that test the effects of mechanisms that are thought to transmit the effects of the hard-to-manipulate variables.[41] Given the rapid pace of innovation, the potential for experimental inquiry remains an open question.

## 1.6 Naturally Occurring Experiments and Quasi-Experiments

Another way to expand the domain of what may be studied experimentally is to seize on *naturally occurring experiments*. Experimental research opportunities arise when interventions are assigned by a government or institution.[42] For example, the

36 Michalopoulos 2005; Green and Gerber 2008.

37 See, e.g., Robins 1985 on income taxes; Newhouse 1989 on health insurance; Krueger and Whitmore 2001 and U.S. Department of Health and Human Services 2010 on schooling. On public housing, see Sanbonmatsu et al. 2006; Harcourt and Ludwig 2006; and Kling, Liebman, and Katz 2007.

38 Olken 2007.

39 Humphreys and Weinstein 2010; Grose 2009.

40 Butler and Nickerson 2011.

41 Ludwig, Kling, and Mullainathan 2011; Card, Della Vigna, and Malmendier 2011.

42 Unfortunately, the term "natural experiment" is sometimes used quite loosely, encompassing not only naturally occurring randomized experiments but also any observational study in which the method of assignment is haphazard or inscrutable. We categorize studies that use near-random or arguably random assignment as quasi-experiments. For definitions of the term *natural experiment* that do not require random assignment, see Dunning 2012 and Shadish, Cook, and Campbell 2002, p. 17.

Vietnam draft lottery,[43] the random assignment of defendants to judges,[44] the random audit of local municipalities in Brazil,[45] lotteries that assign parents the opportunity to place their children in different public schools,[46] the assignment of Indian local governments to be headed by women or members of scheduled castes,[47] the allocation of visas to those seeking to immigrate,[48] and legislative lotteries to determine which representative will be allowed to propose legislation[49] are a few examples where randomization procedures have been employed by government, setting the stage for an experimental analysis. Researchers have also seized on natural experiments conducted by nongovernmental institutions. Universities, for example, occasionally randomize the pairing of roommates, allocation of instructors, and composition of tenure review committees.[50] Sports of all kinds use coin flips and lotteries to assign everything from the sequence of play to the colors worn by the contestants.[51] This list of naturally occurring experimental opportunities might also include revisiting random allocations conducted for other research purposes. A *downstream experiment* refers to a study whose intervention affects not only the proximal outcome of interest but, in so doing, potentially influences other outcomes as well (see Chapter 6). For example, a researcher might revisit an experiment that induced an increase in high school graduation rates in order to assess whether this randomly induced change in educational attainment in turn caused an increase in voter turnout.[52] In this book, we scarcely distinguish between field experiments and naturally occurring experiments, except to note that extra effort is sometimes required in order to verify that draft boards, court systems, or school districts implemented random assignment.

Quite different are *quasi-experiments*, in which near-random processes cause places, groups, or individuals to receive different treatments. Since the mid-1990s, a growing number of scholars have studied instances where institutional rules cause near-random treatment assignments to be allocated among those who fall just short of or just beyond a cutoff, creating a discontinuity. One of the most famous examples of this research design is a study of U.S. congressional districts in which one party's candidate narrowly wins a plurality of votes.[53] The small shift in votes that separates a narrow victory from a narrow defeat produces a treatment—winning the seat in the House of Representatives—that might be construed as random. One

43 Angrist 1991.
44 Kling 2006; Green and Winik 2010.
45 Ferraz and Finan 2008.
46 Hastings, Kane, Staiger, and Weinstein 2007.
47 Beaman et al. 2009; Chattopadhyay and Duflo 2004.
48 Gibson, McKenzie, and Stillman 2011.
49 Loewen, Koop, Settle, and Fowler 2010.
50 Sacerdote 2001; Carrell and West 2010; De Paola 2009; Zinovyeva and Bagues 2010.
51 Hill and Barton 2005; see also Rowe, Harris, and Roberts 2005 for a response to Hill and Barton.
52 Sondheimer and Green 2009.
53 Lee 2008.



could compare near-winners to near-losers in order to assess the effect of a narrow victory on the probability that the winning party wins reelection in the district two years later.

Because quasi-experiments do not involve an explicit random assignment procedure, the causal inferences they support are subject to greater uncertainty. Although the researcher may have good reason to believe that observations on opposite sides of an arbitrary threshold are comparable, there is always some risk that the observations may have "sorted" themselves so as to receive or avoid the treatment. Critics who have looked closely at the pool of congressional candidates who narrowly win or lose have pointed out that there appear to be systematic differences between near-winners and near-losers in terms of their political resources.[54]

The same concerns apply to a wide array of quasi-experiments that take weather patterns, natural disasters, colonial settlement patterns, national boundaries, election cycles, assassinations and so forth to be near-random "treatments." In the absence of random assignment, there is always some uncertainty about how nearly random these treatments are. Although these studies are similar in spirit to field experimentation insofar as they strive to illuminate causal effects in real-world settings, they fall outside the scope of this book because they rely on argumentation rather than randomization procedures. In order to present a single, coherent perspective on experimental design and analysis, this book confines its attention to randomized experiments.

## 1.7 Plan of the Book

This chapter has introduced a variety of important concepts without pausing for rigorous definitions or proofs. Chapter 2 delves more deeply into the properties of experiments, describing in detail the underlying assumptions that must be met for experiments to be informative. Chapter 3 introduces the concept of sampling variability, the statistical uncertainty introduced whenever subjects are randomly allocated to treatment and control groups. Chapter 4 focuses on how covariates, variables that are measured prior to the administration of the treatment, may be used in

54 Grimmer et al. 2011; Caughey and Sekhon 2011. In addition, regression discontinuity analyses often confront the following conundrum: the causal effect of the treatment is identified at the point of discontinuity, but data are sparse in the close vicinity of the boundary. One may expand the comparison to include observations farther from the boundary, but doing so jeopardizes the comparability of groups that do or do not receive the treatment. In an effort to correct for unmeasured differences between the groups, researchers typically use regression to control for trends on either side of the boundary, a method that introduces a variety of modeling decisions and attendant uncertainty. See Imbens and Lemieux 2008 and Green et al. 2009.

experimental design and analysis. Chapters 5 and 6 discuss the complications that arise when subjects are assigned one treatment but receive another. The so-called *noncompliance* or *failure-to-treat* problem is sufficiently common and conceptually challenging to warrant two chapters. Chapter 7 addresses the problem of attrition, or the failure to obtain outcome measurements for every subject. Because field experiments are frequently conducted in settings where subjects communicate, compare, or remember treatments, Chapter 8 considers the complications associated with interference between experimental units. Because researchers are often interested in learning about the conditions under which treatment effects are especially strong or weak, Chapter 9 discusses the detection of heterogeneous treatment effects. Chapter 10 considers the challenge of studying the causal pathways by which an experimental effect is transmitted. Chapter 11 discusses how one might draw generalizations that go beyond the average treatment effect observed in a particular sample and apply them to the average treatment effect in a broader population. The chapter provides a brief introduction to meta-analysis, a statistical technique that pools data from multiple experiments in order to summarize the findings of a research literature. Chapter 12 discusses a series of noteworthy experiments in order to highlight important principles introduced in previous chapters. Chapter 13 guides the reader through the composition of an experimental research report, providing a checklist of key aspects of any experiment that must be described in detail. Appendix A discusses regulations that apply to research involving human subjects. In order to encourage you to put the book's ideas to work, Appendix B suggests several experimental projects that involve low cost and minimal risk to human subjects.

## SUGGESTED READINGS

Accessible introductions to experimental design in real-world settings can be found in Shadish, Cook, and Campbell 2002 and Torgerson and Torgerson 2008. For a discussion of the limitations of field experimentation, see Heckman and Smith 1995. Morgan and Winship (2007), Angrist and Pischke (2009), and Rosenbaum (2010) discuss the challenges of extracting causal inferences from nonexperimental data. Imbens and Lemieux (2008) provide a useful introduction to regression-discontinuity designs.

## EXERCISES: CHAPTER 1

1. Core concepts:
   (a) What is an experiment, and how does it differ from an observational study?
   (b) What is "unobserved heterogeneity," and what are its consequences for the interpretation of correlations?
2. Would you classify the study described in the following abstract as a field experiment, a naturally occurring experiment, a quasi-experiment, or none of the above? Why?

   "This study seeks to estimate the health effects of sanitary drinking water among low-income villages in Guatemala. A random sample of all villages with fewer than 2,000 inhabitants was selected for analysis. Of the 250 villages sampled, 110 were found to have unsanitary drinking water. In these 110 villages, infant mortality rates were, on average, 25 deaths per 1,000 live births, as compared to 5 deaths per 1,000 live births in the 140 villages with sanitary drinking water. Unsanitary drinking water appears to be a major contributor to infant mortality."



3. Based on what you are able to infer from the following abstract, to what extent does the study described seem to fulfill the criteria for a field experiment?

   "We study the demand for household water connections in urban Morocco, and the effect of such connections on household welfare. In the northern city of Tangiers, among homeowners without a private connection to the city's water grid, a random subset was offered a simplified procedure to purchase a household connection on credit (at a zero percent interest rate). Take-up was high, at 69%. Because all households in our sample had access to the water grid through free public taps . . . household connections did not lead to any improvement in the quality of the water households consumed; and despite a significant increase in the quantity of water consumed, we find no change in the incidence of waterborne illnesses. Nevertheless, we find that households are willing to pay a substantial amount of money to have a private tap at home. Being connected generates important time gains, which are used for leisure and social activities, rather than productive activities."[55]

4. A parody appearing in the *British Medical Journal* questioned whether parachutes are in fact effective in preventing death when skydivers are presented with severe "gravitational challenge."[56] The authors point out that no randomized trials have assigned parachutes to skydivers. Why is it reasonable to believe that parachutes are effective even in the absence of randomized experiments that establish their efficacy?

55 Devoto et al. 2011.
56 Smith and Pell 2003.

# CHAPTER 2

# Causal Inference and Experimentation

Although the logic of experimentation is for the most part intuitive, researchers can run into trouble if they lack a firm grasp of the key assumptions that must be met in order for experiments to provide reliable assessments of cause and effect. This point applies in particular to field experimental researchers, who must frequently make real-time decisions about research design. Failure to understand core statistical principles and their practical implications may cause researchers to squander resources and experimental opportunities. It is wise, therefore, to invest time studying the formal statistical properties of experiments before launching a research project.

This chapter introduces a system of notation that will be used throughout the book. By depicting the outcomes that potentially manifest themselves depending on whether the treatment is administered to each unit, the notation clarifies a number of key concepts, such as the idea of a treatment effect. This notational system is then used to shed light on the conditions under which experiments provide persuasive evidence about cause and effect. The chapter culminates with a list of core assumptions and what they imply for experimental design. The advantage of working methodically from core principles is that a long list of design-related admonitions flows from a relatively compact set of ideas that can be stored in working memory.

## 2.1 Potential Outcomes

Suppose we seek to gauge the causal effect of a treatment. For concreteness, suppose we wish to study the budgetary consequences of having women, rather than men, head Indian village councils, which govern rural areas in West Bengal and Rajasthan.[1]

1 See Chattopadhyay and Duflo 2004.

**What you will learn from this chapter:**

1. The system of notation used to describe potential outcomes.
2. Definitions of core terms: average treatment effect, expectation, random assignment, and unbiasedness.
3. Assumptions that must be met in order for experiments to produce unbiased estimates of the average treatment effect.

Students of legislative politics have argued that women bring different policy priorities to the budgetary process in developing countries, emphasizing health issues such as providing clean drinking water. Leave aside for the time being the question of how this topic might be studied using randomly assigned treatments. For the moment, simply assume that each village either receives the treatment (a woman serves as village council head) or remains untreated (with its village council headed by a man). For each village, we also observe the share of the local council budget that is allocated to providing clean drinking water. To summarize, we observe the treatment (whether the village head is a woman or not) and the outcome (what share of the budget goes to a policy issue of special importance to women).

What we do not observe is how the budget in each village headed by a man would have been allocated if it had been headed by a woman, and vice versa. Although we do not observe these counterfactual outcomes, we can nevertheless imagine them. Taking this mental exercise one step further, we might imagine that each village has two *potential outcomes*: the budget it would enact if headed by a woman and the budget it would enact if headed by a man. The gender of the village head determines which potential budget we observe. The other budget remains imaginary or counterfactual.

Table 2.1 provides a stylized example of seven villages in order to introduce the notation that we will use throughout the book. The villages constitute the subjects in this experiment. Each subject is identified by a subscript $i$, which ranges from 1 to 7. The third village on the list, for example, would be designated as $i = 3$. The table imagines what would happen under two different scenarios. Let $Y_i(1)$ be the outcome if village $i$ is exposed to the treatment (a woman as village head), and let $Y_i(0)$ be the outcome if this village is not exposed to the treatment. For example, Village 3 allocates 30% of its budget to water sanitation if headed by a woman but only 20% if headed by a man, so, $Y_3(1) = 30\%$, and $Y_3(0) = 20\%$. These are called potential outcomes because they describe what would happen if a treatment were or were not administered.

For purposes of this example, we assume that each village has just two potential outcomes, depending on whether it receives the treatment; villages are assumed to be unaffected by the treatments that other villages receive. In section 2.7, we spell out more precisely the assumptions that underlie the model of potential outcomes and discuss complications that arise when subjects are affected by the treatments that other subjects receive.



**TABLE 2.1**

Illustration of potential outcomes for local budgets when village council heads are women or men. (Entries are shares of local budgets allocated to water sanitation.)

| *Village i* | $Y_i(0)$ *Budget share if village head is male* | $Y_i(1)$ *Budget share if village head is female* | $\tau_i$ *Treatment effect* |
|---|---|---|---|
| Village 1 | 10 | 15 | 5 |
| Village 2 | 15 | 15 | 0 |
| Village 3 | 20 | 30 | 10 |
| Village 4 | 20 | 15 | −5 |
| Village 5 | 10 | 20 | 10 |
| Village 6 | 15 | 15 | 0 |
| Village 7 | 15 | 30 | 15 |
| **Average** | **15** | **20** | **5** |

## 2.2 Average Treatment Effects

For each village, the causal effect of the treatment ($\tau_i$) is defined as the difference between two potential outcomes:

$$\tau_i \equiv Y_i(1) - Y_i(0). \quad [2.1]$$

In other words, the treatment effect for each village is the difference between two potential states of the world, one in which the village receives the treatment and another in which it does not. For Village 3, this causal effect is $30 - 20 = 10$.

The empirical challenge that researchers typically face when observing outcomes is that at any given time one can observe $Y_i(1)$ or $Y_i(0)$ but not both. (Bear in mind that the only reason we are able to see both potential outcomes for each village in Table 2.1 is that this is a hypothetical example!) Building on the notational system introduced above, we define $Y_i$ as the observed outcome in each village and $d_i$ as the observed treatment that is delivered in each village. In this case, $Y_i$ is the observed share of the budget allocated to water sanitation, and $d_i$ equals 1 when a woman is village head and 0 otherwise.

**BOX 2.1**

### Potential Outcomes Notation

In this system of notation, the subscript $i$ refers to subjects 1 through $N$.

The variable $d_i$ indicates whether the $i$th subject is treated: $d_i = 1$ means the $i$th subject receives the treatment, and $d_i = 0$ means the $i$th subject does not receive the treatment. It is assumed that $d_i$ is observed for every subject.

$Y_i(1)$ is the potential outcome if the $i$th subject were treated. $Y_i(0)$ is the potential outcome if the $i$th subject were not treated. In general, potential outcomes may be written $Y_i(d)$, where $d$ indexes the treatment. These potential outcomes are fixed attributes of each subject and represent the outcome that would be observed hypothetically if that subject were treated or untreated.

A schedule of potential outcomes refers to a comprehensive list of potential outcomes for all subjects. The rows of this schedule are indexed by $i$, and the columns are indexed by $d$. For example, in Table 2.1 the $Y_i(0)$ and $Y_i(1)$ potential outcomes for the fifth subject may be found in adjacent columns of the fifth row.

The connection between the observed outcome $Y_i$ and the underlying potential outcomes is given by the equation $Y_i = d_iY_i(1) + (1 - d_i)Y_i(0)$. This equation indicates that the $Y_i(1)$ are observed for subjects who are treated, and the $Y_i(0)$ are observed for subjects who are not treated. For any given subject, we observe either $Y_i(1)$ or $Y_i(0)$, never both.

It is sometimes useful to refer to potential outcomes for a subset of all subjects. Expressions of the form $Y_i(\cdot) \mid X = x$ denote potential outcomes when the condition $X = x$ holds. For example, $Y_i(0) \mid d_i = 1$ refers to the untreated potential outcome for a subject who actually receives the treatment.

Because we often want to know about the statistical properties of a hypothetical random assignment, we distinguish between $d_i$, the treatment that a given subject receives (a variable that one observes in an actual dataset), and $D_i$, the treatment that could be administered hypothetically. $D_i$ is a random variable, and the $i$th subject might be treated in one hypothetical study and not in another. For example, $Y_i(1) \mid D_i = 1$ refers to the treated potential outcome for a subject who would be treated under some hypothetical allocation of treatments.



The budget that we observe in each village may be summarized using the following expression:

$$Y_i = d_iY_i(1) + (1 - d_i)Y_i(0). \tag{2.2}$$

Because $d_i$ is either 0 or 1, one of the terms on the right side of the equals sign will always be zero. We observe the potential outcome that results from treatment, $Y_i(1)$, if the treatment is administered ($d_i = 1$). If the treatment is not administered ($d_i = 0$), we observe the potential outcome that results when no treatment occurs, $Y_i(0)$.

The average treatment effect, or ATE, is defined as the sum of the $\tau_i$ divided by $N$, the number of subjects:

$$\text{ATE} \equiv \frac{1}{N}\sum_{i=1}^{N}\tau_i. \tag{2.3}$$

An equivalent way to obtain the average treatment effect is to subtract the average value of $Y_i(0)$ from the average value of $Y_i(1)$:

$$\frac{1}{N}\sum_{i=1}^{N} Y_i(1) - \frac{1}{N}\sum_{i=1}^{N} Y_i(0) = \frac{1}{N}\sum_{i=1}^{N}(Y_i(1) - Y_i(0)) = \frac{1}{N}\sum_{i=1}^{N}\tau_i. \tag{2.4}$$

The average treatment effect is an extremely important concept. Villages may have different $\tau_i$, but the ATE indicates how outcomes would change on average if every village were to go from untreated (male village council head) to treated (female village council head).

From the rightmost column of Table 2.1, we can calculate the ATE for the seven villages. The average treatment effect in this example is 5 percentage points: if all villages were headed by men, they would on average spend 15% of their budgets on water sanitation, whereas if all villages were headed by women, this figure would rise to 20%.

**BOX 2.2**

### Definition: Average Treatment Effect

The average treatment effect (ATE) is the sum of the subject-level treatment effects, $Y_i(1) - Y_i(0)$, divided by the total number of subjects. An equivalent way to express the ATE is to say that it equals $\mu_{Y(1)} - \mu_{Y(0)}$, where $\mu_{Y(1)}$ is the average value of $Y_i(1)$ for all subjects and $\mu_{Y(0)}$ is the average value of $Y_i(0)$ for all subjects.

## 2.3 Random Sampling and Expectations

Suppose that instead of calculating the average potential outcome for all villages, we drew a random sample of villages and calculated the average among the villages we sampled. By *random sample*, we mean a selection procedure in which $v$ villages are selected from the list of $N$ villages, and every possible set of $v$ villages is equally likely to be selected. For example, if we select one village at random from a list of seven villages, seven possible samples are equally likely. If we select three villages at random from a list of seven villages,

$$\frac{N!}{v!(N-v)!} = \frac{7!}{3!4!} = \frac{7\times6\times5\times4\times3\times2\times1}{(3\times2\times1)(4\times3\times2\times1)} = 35 \quad [2.5]$$

possible samples are equally likely. If potential outcomes vary from one village to the next, the average potential outcome in the villages we sample will vary, depending on which of the possible samples we happen to select. The sample average may be characterized as a *random variable*, a quantity that varies from sample to sample.

The term *expected value* refers to the average outcome of a random variable. (See Box 2.3.) In our example, the random variable is the number we obtain when we sample villages at random and calculate their average outcome. Recall from introductory statistics that under random sampling, the expected value of a sample average is equal to the average of the population from which the sample is drawn.[2] This principle may be illustrated using the population of villages depicted in Table 2.1. Recall that the average value of $Y_i(0)$ among all villages in Table 2.1 is 15. Suppose we sample two villages at random from the list of seven villages and calculate the average value of $Y_i(0)$ for the two selected villages. There are

$$\frac{N!}{v!(N-v)!} = \frac{7!}{2!5!} = 21 \quad [2.6]$$

possible ways of sampling two villages at random from a list of seven, and each sample is equally likely to be drawn. Any given sample of two villages might contain an average value of $Y_i(0)$ that is higher or lower than the true average of 15, but the expected value refers to what we would obtain on average if we were to examine all 21 possible samples, for each one calculating the average value of $Y_i(0)$:

$$\{10, 12.5, 12.5, 12.5, 12.5, 12.5, 12.5, 15, 15, 15, 15, 15, 15, 15, 17.5, 17.5, 17.5, 17.5, 17.5, 17.5, 20\}. \quad [2.7]$$

2 The easiest way to see the intuition behind this principle is to consider the case in which we randomly sample just one village. Each village is equally likely to be sampled. The average over all seven possible samples is identical to the average for the entire population of seven villages. This logic generalizes to samples where $v > 1$ because each village appears in exactly $v/7$ of all possible samples.



**BOX 2.3**

The expectation of a discrete random variable $X$ is defined as

$$E[X] = \sum x \Pr[X = x],$$

where $\Pr[X = x]$ denotes the probability that $X$ takes on the value $x$, and where the summation is taken over all possible values of $x$.

For example, what is the expected value of a randomly selected value of $\tau_i$ from Table 2.1?

$$E[\tau_i] = \sum \tau \Pr[\tau_i = \tau]$$
$$= (-5)\left(\frac{1}{7}\right) + (0)\left(\frac{2}{7}\right) + (5)\left(\frac{1}{7}\right) + (10)\left(\frac{2}{7}\right) + (15)\left(\frac{1}{7}\right) = 5.$$

**Properties of Expectations**

The expectation of the constant $\alpha$ is itself: $E[\alpha] = \alpha$.

For a random variable $X$ and constants $\alpha$ and $\beta$, $E[\alpha + \beta X] = \alpha + \beta E[X]$.

The expectation of a sum of two random variables, $X$ and $Y$, is the sum of their expectations: $E[X + Y] = E[X] + E[Y]$.

The expectation of the product of two random variables, $X$ and $Y$, is the product of their expectations plus the covariance between them: $E[XY] = E[X]E[Y] + E[(X - E[X])(Y - E[Y])]$.

The average of these 21 numbers is 15. In other words, the expected value of the average $Y_i(0)$ obtained from a random sample of two villages is 15.

The concept of expectations plays an important role in the discussion that follows. Because we will refer to expectations so often, a bit more notation is helpful. The notation $E[X]$ refers to the expectation of a random variable $X$. (See Box 2.3.) The expression "the expected value of $Y_i(0)$ when one subject is sampled at random" will be written compactly as $E[Y_i(0)]$. When a term like $Y_i(0)$ appears in conjunction with an expectations operator, it should be read not as the value of $Y_i(0)$ for subject $i$ but instead as a random variable that is equal to the value of $Y_i(0)$ for a randomly selected subject. When the expression $E[Y_i(0)]$ is applied to values in Table 2.1, the random variable is the random selection of a $Y_i(0)$ from the list of all $Y_i(0)$; since there are seven possible random selections, the average of which is 15, it follows that $E[Y_i(0)] = 15$.

Sometimes attention is focused on the expected value of a random variable within a subgroup. *Conditional expectations* refer to subgroup averages. In terms of notation, the logical conditions following the | symbol indicate the criteria that define the subgroup. For example, the expression "the expectation of $Y_i(1)$ when one village is selected at random from those villages that were treated" is written $E[Y_i(1) \mid d_i = 1]$. The idea of a conditional expectation is straightforward when working with quantities that are in principle observable. More mind-bending are expressions like $E[Y_i(1) \mid d_i = 0]$, which denotes "the expectation of $Y_i(1)$ when one village is selected at random from those villages that were not treated." In the course of conducting research, we will never actually see $Y_i(1)$ for an untreated village, nor will we see $Y_i(0)$ for a treated village. These potential outcomes can be imagined but not observed.

One special type of conditional expectation arises when the subgroup is defined by the outcome of a random process. In that case, the conditional expectation may vary depending on which subjects happened to meet the condition in any particular realization of the random process. For example, suppose that a random process, such as a coin flip, determines which subjects are treated. For a given treatment assignment $d_i$, we could calculate $E[Y_i(1) \mid d_i = 0]$, but this expectation might have been different had the coin flips come out differently. Suppose we want to know the expected conditional expectation, or how the conditional expectation would come out, on average, across all possible ways that $d_i$ could have been allocated. Let $D_i$ be a random variable that indicates whether each subject would be treated in a hypothetical experiment. The conditional expectation $E[Y_i(1) \mid D_i = 0]$ is calculated by considering all possible realizations of $D_i$ (all the possible ways that $N$ coins could have been flipped) in order to form the joint probability distribution function for $Y_i(1)$ and $D_i$. As long as we know the joint probability of observing each paired set of values $\{Y(1), D\}$, we can calculate the conditional expectation using the formula in Box 2.4.[3]

With this basic system of notation in place, we may now describe the connection between expected potential outcomes and the average treatment effect (ATE):

$$\begin{aligned} E[Y_i(1) - Y_i(0)] &= E[Y_i(1)] - E[Y_i(0)] \\ &= \frac{1}{N}\sum_{i=1}^{N} Y_i(1) - \frac{1}{N}\sum_{i=1}^{N} Y_i(0) \\ &= \frac{1}{N}\sum_{i=1}^{N} [Y_i(1) - Y_i(0)] \equiv \text{ATE}. \end{aligned} \tag{2.8}$$

3 The notation $E[Y_i(1) \mid D_i = 0]$ may be regarded as shorthand for $E[E[Y_i(1) \mid d_i = 0, \boldsymbol{d}]]$, where $\boldsymbol{d}$ refers to a vector of treatment assignments and $d_i$ refers its $i$th element. Given $\boldsymbol{d}$, we may calculate the probability distribution function for all $\{Y(1), d\}$ pairs and the expectation given this set of assignments. Then we may take the expectation of this expected value by summing over all possible $\boldsymbol{d}$ vectors.



**BOX 2.4**

**Definition: Conditional Expectation**

For discrete random variables $Y$ and $X$, the conditional expectation of $Y$ given that $X$ takes on the value $x$ is

$$E[Y \mid X = x] = \sum y \Pr[Y = y \mid X = x] = \sum y \frac{\Pr[Y = y, X = x]}{\Pr[X = x]},$$

where $\Pr[Y = y, X = x]$ denotes the joint probability of $Y = y$ and $X = x$, and where the summation is taken over all possible values of $y$.

For example, in Table 2.1 what is the conditional expectation of a randomly selected value of $\tau_i$, for villages where $Y_i(0) > 10$? This question requires us to describe the joint probability distribution function for the variables $\tau_i$ and $Y_i(0)$ so that we can calculate $\Pr[\tau_i = \tau, Y_i(0) > 10]$. Table 2.1 indicates that the $\{\tau, Y(0)\}$ pair $\{0, 15\}$ occurs with probability $2/7$, while the other pairs $\{5, 10\}$, $\{10, 20\}$, $\{-5, 20\}$, $\{10, 10\}$, and $\{15, 15\}$ each occur with probability $1/7$. The marginal distribution of $Y_i(0)$ reveals that 5 of the 7 $Y_i(0)$ are greater than 10, so $\Pr[Y_i(0) > 10] = 5/7$.

$$\begin{aligned} E[\tau_i \mid Y_i(0) > 10] &= \sum \tau \frac{\Pr[\tau_i = \tau, Y_i(0) > 10]}{\Pr[Y_i(0) > 10]} \\ &= (-5)\frac{\frac{1}{7}}{\frac{5}{7}} + (0)\frac{\frac{2}{7}}{\frac{5}{7}} + (5)\frac{0}{\frac{5}{7}} + (10)\frac{\frac{1}{7}}{\frac{5}{7}} + (15)\frac{\frac{1}{7}}{\frac{5}{7}} = 4. \end{aligned}$$

In order to illustrate the idea of a conditional expectation when conditioning on the outcome of a random process, suppose we randomly assign one of the observations in Table 2.1 to treatment ($D_i = 1$) and the remaining six observations to control ($D_i = 0$). If each of the seven possible assignments occurs with probability $1/7$, what is the expected value of a randomly selected $\tau_i$ given that $D_i = 1$? Again, we start with the joint probability density function for $\tau_i$ and $D_i$ and consider all possible pairings of these two variables' values. The $\{\tau, D\}$ pairings $\{-5, 1\}$, $\{5, 1\}$, and $\{15, 1\}$ occur with probability $1/49$, while the pairings $\{0, 1\}$ and $\{10, 1\}$ occur with probability $2/49$; the remaining $\{\tau, D\}$ pairings are instances in which $\tau$ is paired with 0. The marginal distribution $\Pr[D_i = 1] = 3(1/49) + 2(2/49) = 1/7$.

$$\begin{aligned} E[\tau_i \mid D_i = 1] &= \sum \tau \frac{\Pr[\tau_i = \tau, D_i = 1]}{\Pr[D_i = 1]} \\ &= (-5)\frac{\frac{1}{49}}{\frac{1}{7}} + (0)\frac{\frac{2}{49}}{\frac{1}{7}} + (5)\frac{\frac{1}{49}}{\frac{1}{7}} + (10)\frac{\frac{2}{49}}{\frac{1}{7}} + (15)\frac{\frac{1}{49}}{\frac{1}{7}} = 5. \end{aligned}$$

The first line of equation (2.8) expresses the fact that when a village is selected at random from the list of villages, its expected treatment effect is equal to the difference between the expected value of a randomly selected treated potential outcome and the expected value of a randomly selected untreated potential outcome. The second equality in equation (2.8) indicates that the expected value of a randomly selected $Y_i(1)$ equals the average of all $Y_i(1)$ values, and that the expected value of a randomly selected $Y_i(0)$ equals the average of all $Y_i(0)$ values. The third equality reflects the fact that the difference between the two averages in the second line of equation (2.8) can be expressed as the average difference in potential outcomes. The final equality notes that the average difference in potential outcomes is the definition of the average treatment effect. In sum, the difference in expectations equals the difference in average potential outcomes for the entire list of villages, or the ATE.[4]

This relationship is apparent from the schedule of potential outcomes in Table 2.1. The column of numbers representing the treatment effect ($\tau_i$) is, on average, 5. If we were to select villages at random from this list, we would expect their average treatment effect to be 5. We get the same result if we subtract the expected value of a randomly selected $Y_i(0)$ from the expected value of a randomly selected $Y_i(1)$.

## 2.4 Random Assignment and Unbiased Inference

The challenge of estimating the average treatment effect is that at a given point in time each village is either treated or not: either $Y_i(1)$ or $Y_i(0)$ is observed, but not both. To illustrate the problem, Table 2.2 shows what outcomes would be observed if Village 1 and Village 7 were treated, while the remaining villages were not. We observe $Y_i(1)$ for Villages 1 and 7 but not $Y_i(0)$. For Villages 2, 3, 4, 5, and 6, we observe $Y_i(0)$ but not $Y_i(1)$. The unobserved or "missing" values in Table 2.2 are indicated with a "?".

4 The notation used here is just one way to explicate the link between expectations and the ATE. Samii and Aronow (2012) suggest an alternative formalization. Their model envisions a finite population $U$ consisting of units $j$ in 1, 2, . . . , $N$, each of which has an associated triple $(y_j(1), y_j(0), D'_j)$ such that $y_j(1)$ and $y_j(0)$ are fixed potential outcomes and $D'_j$ is a random variable indicating the treatment status of unit $j$. Reassign a random index ordering $i$ in 1, 2, . . . , $N$. Then, for an arbitrary unit $i$, there exists an associated triple of random variables $(Y_i(1), Y_i(0), D_i)$ such that the random variable $Y_i = D_iY_i(1) + (1 - D_i)Y_i(0)$. It follows that for equation (2.8):

$$E[Y_i(1)] - E[Y_i(0)] = \frac{1}{N}\sum_{j=1}^{N} y_j(1) - \frac{1}{N}\sum_{j=1}^{N} y_j(0) = \text{ATE}.$$

Statistical operators such as expectations or independence refer to random variables associated with an arbitrary index $i$. Looking ahead to later chapters, one might expand this system to include other unit-level attributes, such as covariates or missingness, by attaching them to the triple indexed by $j$ before reassigning the ordering.



**TABLE 2.2**

Illustration of observed outcomes for local budgets when two village councils are headed by women.

| *Village i* | $Y_i(0)$ *Budget share if village head is male* | $Y_i(1)$ *Budget share if village head is female* | $\tau_i$ *Treatment effect* |
|---|---|---|---|
| Village 1 | ? | 15 | ? |
| Village 2 | 15 | ? | ? |
| Village 3 | 20 | ? | ? |
| Village 4 | 20 | ? | ? |
| Village 5 | 10 | ? | ? |
| Village 6 | 15 | ? | ? |
| Village 7 | ? | 30 | ? |
| **Estimated average based on observed data** | **16** | **22.5** | **6.5** |

*Note:* The observed outcomes in this table are based on the potential outcomes listed in Table 2.1.

Random assignment addresses the "missing data" problem by creating two groups of observations that are, in expectation, identical prior to application of the treatment. When treatments are allocated randomly, the treatment group is a random sample of all villages, and therefore the expected potential outcomes among villages in the treatment group are identical to the average potential outcomes among all villages. The same is true for villages in the control group. The control group's expected potential outcomes are also identical to the average potential outcomes among all villages. Therefore, in expectation, the treatment group's potential outcomes are the same as the control group's. Although any given random allocation of villages to treatment and control groups may produce groups of villages that have different average potential outcomes, this procedure is fair in the sense that it does not tend to give one group a higher set of potential outcomes than the other.

As Chattopadhyay and Duflo point out, random assignment is in fact used in rural India to assign women to head one-third of the local village councils.[5] Ordinarily, men would head the village councils, but Indian law mandates that selected

5 Chattopadhyay and Duflo 2004. A lottery is used to assign council positions to women in Rajasthan. In West Bengal, a near-random assignment procedure is used whereby villagers are assigned according to their serial numbers.

villages install a female representative as head of the council. For purposes of illustration, suppose that our collection of seven villages were subject to this law, and that two villages will be randomly assigned female council heads. Consider the statistical implications of this arrangement. This random assignment procedure implies that every village has the same probability of receiving the treatment; assignment bears no systematic relationship to villages' observed or unobserved attributes.

Let's take a closer look at the formal implications of this form of random assignment. When villages are assigned such that every village has the same probability of receiving the treatment, the villages that are randomly chosen for treatment are a random subset of the entire set of villages. Therefore, the expected $Y_i(1)$ potential outcome among treated villages is the same as the expected $Y_i(1)$ potential outcome for the entire set of villages:

$$E[Y_i(1) \mid D_i = 1] = E[Y_i(1)]. \tag{2.9}$$

**BOX 2.5**

**Two Commonly Used Forms of Random Assignment**

Random assignment refers to a procedure that allocates treatments with known probabilities that are greater than zero and less than one.

The most basic forms of random assignment allocate treatments such that every subject has the same probability of being treated. Let $N$ be the number of subjects, and let $m$ be the number of subjects who are assigned to the treatment group. Assume that $N$ and $m$ are integers such that $0 < m < N$. Simple random assignment refers to a procedure whereby each subject is allocated to the treatment group with probability $m/N$. Complete random assignment refers to a procedure that allocates exactly $m$ units to treatment.

Under simple or complete random assignment, the probability of being assigned to the treatment group is identical for all subjects; therefore treatment status is statistically independent of the subjects' potential outcomes and their background attributes ($\boldsymbol{X}$):

$$Y_i(0), Y_i(1), \boldsymbol{X} \perp\!\!\!\perp D_i,$$

where the symbol $\perp\!\!\!\perp$ means "is independent of." For example, if a die roll is used to assign subjects to treatment with probability $1/6$, knowing whether a subject is treated provides no information about the subject's potential outcomes or background attributes. Therefore, the expected value of $Y_i(0)$, $Y_i(1)$, and $X_i$ is the same in treatment and control groups.



When we randomly select villages into the treatment group, the villages we leave behind for the control group are also a random sample of all villages. The expected $Y_i(1)$ in the control group ($D_i = 0$) is therefore equal to the expected $Y_i(1)$ for the entire set of villages:

$$E[Y_i(1) \mid D_i = 0] = E[Y_i(1)]. \tag{2.10}$$

Putting equations (2.9) and (2.10) together, we see that under random assignment the treatment and control groups have the same expected potential outcome:

$$E[Y_i(1) \mid D_i = 1] = E[Y_i(1) \mid D_i = 0]. \tag{2.11}$$

Equation (2.11) also underscores the distinction between realized and unrealized potential outcomes. On the left side of the equation is the expected treated potential outcome among villages that receive the treatment. The treatment causes this potential outcome to become observable. On the right side of the equation is the expected treated potential outcome among villages that do not receive the treatment. Here, the lack of treatment means that the treated potential outcome remains unobserved for these subjects.

The same logic applies to the control group. Villages that do not receive the treatment ($D_i = 0$) have the same expected untreated potential outcome $Y_i(0)$ that the treatment group ($D_i = 1$) would have if it were untreated:

$$E[Y_i(0) \mid D_i = 0] = E[Y_i(0) \mid D_i = 1] = E[Y_i(0)]. \tag{2.12}$$

Equations (2.11) and (2.12) follow from random assignment: $D_i$ conveys no information whatsoever about the potential values of $Y_i(1)$ or $Y_i(0)$. The randomly assigned values of $D_i$ determine which value of $Y_i$ we actually *observe*, but they are nevertheless statistically independent of the *potential* outcomes $Y_i(1)$ and $Y_i(0)$. (See Box 2.5 for discussion of the term *independence*.)

When treatments are assigned randomly, we may rearrange equations (2.8), (2.11), and (2.12) in order to express the average treatment effect as

$$\text{ATE} = E[Y_i(1) \mid D_i = 1] - E[Y_i(0) \mid D_i = 0]. \tag{2.13}$$

This equation suggests an empirical strategy for estimating the average treatment effect. The terms $E[Y_i(1) \mid (D_i = 1)]$ and $E[Y_i(0) \mid (D_i = 0)]$ may be estimated using experimental data. We do not observe the $Y_i(1)$ potential outcomes for all observations, but we do observe them for the random sample of observations that receive the treatment. Similarly, we do not observe the $Y_i(0)$ potential outcomes for all observations, but we do observe them for the random sample of observations in the control group. If we want to estimate the average treatment effect, equation (2.13) suggests that we should take the difference between two sample means: the average

outcome in the treatment group minus the average outcome in the control group. Ideas that enable researchers to use observable quantities (e.g., sample averages) to reveal parameters of interest (e.g., average treatment effects) are termed *identification* strategies.

Statistical procedures used to make guesses about parameters such as the average treatment effect are called *estimators*. In this example, the estimator is very simple, just a difference between two sample averages. Before applying an estimator to actual data, a researcher should reflect on its statistical properties. One especially important property is *unbiasedness*. An estimator is unbiased if it generates the right answer, on average. In other words, if the experiment were replicated an infinite number of times under identical conditions, the average *estimate* would equal the true parameter. Some guesses may be too high and others too low, but the average guess will be correct. In practice, we will not be able to perform an infinite number of experiments. In fact, we might just perform one experiment and leave it at that. Nevertheless, in theory we can analyze the properties of our estimation procedure to see whether, on average, it recovers the right answer. (In the next chapter, we consider another property of estimators: how precisely they estimate the parameter of interest.)

In sum, when treatments are administered using a procedure that gives every subject the same probability of being treated, potential outcomes are independent of the treatments that subjects receive. This property suggests an identification strategy for estimating average treatment effects using experimental data.

The remaining task is to demonstrate that the proposed estimator—the difference between the average outcome in the treatment group and the average outcome in the control group—is an unbiased estimator of the ATE when all subjects have the same probability of being treated. The proof is straightforward. Because the units assigned to the control group are a random sample of all units, the average of the control group outcomes is an unbiased estimator of the average value of $Y_i(0)$

**BOX 2.6**

**Definition: Estimator and Estimate**

An estimator is a procedure or formula for generating guesses about parameters such as the average treatment effect. The guess that an estimator generates based on a particular experiment is called an estimate. Estimates are denoted using a "hat" notation. The estimate of the parameter $\theta$ is written $\hat{\theta}$.



among all units. The same goes for the treatment group: the average outcome among units that receive the treatment is an unbiased estimator of the average value of $Y_i(1)$ among all units. Formally, if we randomly shuffle the villages and place the first $m$ subjects in the treatment group and the remaining $N - m$ subjects in the control group, we can analyze the expected, or average, outcome over all possible random assignments:

$$
\begin{aligned}
E\left[\frac{\sum_1^m Y_i}{m} - \frac{\sum_{m+1}^N Y_i}{N-m}\right] &= E\left[\frac{\sum_1^m Y_i}{m}\right] - E\left[\frac{\sum_{m+1}^N Y_i}{N-m}\right] \\
&\quad \text{Average outcome among treated units} \qquad \text{Average outcome among untreated units} \\
&= \frac{E[Y_1] + E[Y_2] + \cdots + E[Y_m]}{m} - \frac{E[Y_{m+1}] + E[Y_{m+2}] + \cdots + E[Y_N]}{N-m} \\
&= E[Y_i(1) \mid D_i = 1] - E[Y_i(0) \mid D_i = 0] \\
&= E[Y_i(1)] - E[Y_i(0)] = E[\tau_i] = \text{ATE}.
\end{aligned}
\tag{2.14}
$$

Equation (2.14) conveys a simple but extremely useful idea. When units are randomly assigned, a comparison of average outcomes in treatment and control groups (the so-called *difference-in-means estimator*) is an unbiased estimator of the average treatment effect.

**BOX 2.7**

**Definition: Unbiased Estimator**

An estimator is unbiased if the expected value of the estimates it produces is equal to the true parameter of interest. Call $\theta$ the parameter we seek to estimate, such as the ATE. Let $\hat{\theta}$ represent an estimator, or procedure for generating estimates. For example, $\hat{\theta}$ may represent the difference in average outcomes between treatment and control groups. The expected value of this estimator is the average estimate we would obtain if we apply this estimator to all possible realizations of a given experiment or observational study. We say that $\hat{\theta}$ is unbiased if $E(\hat{\theta}) = \theta$; in words, the estimator $\hat{\theta}$ is unbiased if the expected value of this estimator is $\theta$, the parameter of interest. Although unbiasedness is a property of estimators and not estimates, we refer to the estimates generated by an unbiased estimator as "unbiased estimates."

## 2.5 The Mechanics of Random Assignment

The result in equation (2.14) hinges on random assignment, and so it is important to be clear about what constitutes random assignment. *Simple random assignment* is a term of art, referring to a procedure—a die roll or coin toss—that gives each subject an identical probability of being assigned to the treatment group. The practical drawback of simple random assignment is that when $N$ is small, random chance can create a treatment group that is larger or smaller than what the researcher intended. For example, you could flip a coin to assign each of 10 subjects to the treatment condition, but there is only a 24.6% chance of ending up with exactly 5 subjects in treatment and 5 in control. A useful special case of simple random assignment is *complete random assignment*, where exactly $m$ of $N$ units are assigned to the treatment group with equal probability.[6]

The procedure used to conduct complete random assignment can take any of three equivalent forms. Suppose one has $N$ subjects and seeks to assign treatments to $m$ of them. The first method is to select one subject at random, then select another at random from the remaining units, and so forth until you have selected $m$ subjects into the treatment group. A second method is to enumerate all of the possible ways that $m$ subjects may be selected from a list of $N$ subjects, and randomly select one of the possible allocation schemes. A third method is to randomly permute the order of all $N$ subjects and label the first $m$ subjects as the treatment group.[7]

Beware of the fact that *random* is a word that is used loosely in common parlance to refer to procedures that are arbitrary, haphazard, or unplanned. The problem is that arbitrary, haphazard, or unplanned treatments may follow systematic patterns that go unnoticed. Procedures such as alternation are risky because there may be systematic reasons why certain types of subjects might alternate in a sequence, and indeed, some early medical experiments ran into exactly this problem.[8] We use the term *random* in a more exacting sense. The physical or electronic procedure by which randomization is conducted ensures that assignment to the treatment group is statistically independent of all observed or unobserved variables.

6 In Chapters 3 and 4, we discuss other frequently used methods of random assignment: clustered random assignment, where groups of subjects are randomly assigned to treatment and control, and block random assignment (also called stratified random assignment), where individuals are first divided into blocks, and then random assignment is performed within each block. Box 2.5 notes that a defining feature of complete (as opposed to clustered or blocked) random assignment is that all possible assignments of $N$ subjects to a treatment group of size $m$ are equally likely.

7 Cox and Reid 2000, p. 20. The term *complete randomization* is a bit awkward, as the word *complete* does not convey the requirement that exactly $m$ units are allocated to treatment, but this terminology has become standard (see Rosenbaum 2002, pp. 25–26).

8 Hróbjartsson, Gøtzsche, and Gluud 1998.



In practical terms, random assignment is best done using statistical software. Here is an easy procedure for implementing complete random assignment. First, determine $N$, the number of subjects in your experiment, and $m$, the number of subjects who will be allocated to the treatment group. Second, set a random number "seed" using a statistics package, so that your random numbers may be reproduced by anyone who cares to replicate your work. Third, generate a random number for each subject. Fourth, sort the subjects by their random numbers in ascending order. Finally, classify the first $m$ observations as the treatment group. Example programs using R may be found at http://isps.research.yale.edu/FEDAI.

Generating random numbers is just the first step in implementing random assignment. After the numbers are generated, one must take pains to preserve the integrity of the assignment process. A deficiency of alternation and many other arbitrary procedures is that they allow those administering the allocation to foresee who will be assigned to which experimental group. If a receptionist seeks to get the sickest patients into the experimental treatment group and knows that the pattern of assignments alternates, he can reorder the patients in such a way as to shuttle the sickest subjects into the treatment group.[9] The same concern arises even when a random sequence of numbers is used to assign incoming patients: random allocation may be undone if the receptionist knows the order of assignments ahead of time, because that enables him to position patients so that they will be assigned to a certain experimental group. In order to guard against potential threats to the integrity of random assignment, researchers should build extra procedural safeguards into their experimental designs, such as blinding those administering the experiment to the subjects' assigned experimental groups.

## 2.6 The Threat of Selection Bias When Random Assignment Is Not Used

Without random assignment, the identification strategy derived from equation (2.14) unravels. The treatment and control groups are no longer random subsets of all units in the sample. Instead, we confront what is known as a *selection problem*: receiving treatment may be systematically related to potential outcomes. For example, absent random assignment, villages determine whether their councils are headed by women. The villages that end up with female council heads may not be a random subset of all villages.

9 For examples of experiments in which random assignment was subverted, see Torgerson and Torgerson 2008.

To see how nonrandom selection jeopardizes the identification strategy of comparing average outcomes in the treatment and control groups, rewrite the expected difference in outcomes from equation (2.13) by subtracting and adding $E[Y_i(0) \mid D_i = 1]$:

$$\underbrace{E[Y_i(1) \mid D_i = 1] - E[Y_i(0) \mid D_i = 0]}_{\text{Expected difference between treated and untreated outcomes}}$$

$$= \underbrace{E[Y_i(1) - Y_i(0) \mid D_i = 1]}_{\text{ATE among the treated}} + \underbrace{E[Y_i(0) \mid D_i = 1] - E[Y_i(0) \mid D_i = 0]}_{\text{Selection bias}}. \quad (2.15)$$

Under random assignment, the selection bias term is zero, and the ATE among the (randomly) treated villages is the same as the ATE among all villages. In the absence of random assignment, equation (2.15) warns that the apparent treatment effect is a mixture of selection bias and the ATE for a subset of villages.

In order to appreciate the implications of equation (2.15), consider the following scenario. Suppose that instead of randomly selecting villages to receive the treatment, our procedure were to let villages decide whether to take the treatment. Refer back to Table 2.1 and imagine that, if left to their own devices, Village 5 and Village 7 always elect a woman due to villagers' pent-up demand for water sanitation, while the remaining villages always elect a man.[10] Self-selection in this case leads to an exaggerated estimate of the ATE because receiving the treatment is associated with lower-than-average values of $Y_i(0)$ and higher-than-average values of $Y_i(1)$. The average outcome in the treatment group is 25, and the average outcome in the control group is 16. The estimated ATE is therefore 9, whereas the actual ATE is 5. Referring to equation (2.15) we see that in this case the ATE among the treated is not equal to the ATE for the entire subject pool, nor is the selection bias term equal to zero. The broader point is that it is risky to compare villages that choose to receive the treatment with villages that choose not to. In this example, self-selection is related to potential outcomes; as a result, the comparison of treated and untreated villages recovers neither the ATE for the sample as a whole nor the ATE among those villages that receive treatment.

The beauty of experimentation is that the randomization procedure generates a schedule of treatment and control assignments that are statistically independent of

10 When taking expectations over hypothetical replications of an experiment, we consider all possible random assignments. In our example of non-random allocation, however, nature makes the assignment. When taking expectations, we must therefore consider the average of all possible natural assignments. Rather than make up an assortment of possible assignments and stipulate the probability that each scenario occurs, we have kept the example as simple as possible and assumed that the villages "always" elect the same type of candidate. In effect, we are taking expectations over just one possible assignment that occurs with probability 1.



potential outcomes. In other words, the assumptions underlying equations (2.9) to (2.13) are justified by reference to the *procedure* of random assignment, not substantive arguments about the comparability of potential outcomes in the treatment and control groups.

The preceding discussion should not be taken to imply that experimentation invokes no substantive assumptions. The unbiasedness of the difference-in-means estimator hinges not only on random assignment but also on two assumptions about potential outcomes, the plausibility of which will vary depending on the application. The next section spells out these important assumptions.

## 2.7 Two Core Assumptions about Potential Outcomes

To this point, our characterization of potential outcomes has glossed over two important details. In order to ease readers into the framework of potential outcomes, we simply stipulated that each subject has two potential outcomes, $Y_i(1)$ if treated and $Y_i(0)$ if not treated. To be more precise, each potential outcome depends *solely* on whether the subject *itself* receives the treatment. When writing potential outcomes in this way, we are assuming that potential outcomes respond only to the treatment and not some other feature of the experiment, such as the way the experimenter assigns treatments or measures outcomes. Furthermore, potential outcomes are defined over the set of treatments that the subject itself receives, not the treatments assigned to other subjects. In technical parlance, the "solely" assumption is termed *excludability* and the "itself" assumption is termed *non-interference*.

### 2.7.1 Excludability

When we define two, and only two, potential outcomes based on whether the treatment is administered, we implicitly assume that the only relevant causal agent is receipt of the treatment. Because the point of an experiment is to isolate the causal effect of the treatment, our schedule of potential outcomes excludes from consideration factors other than the treatment. When conducting an experiment, therefore, we must define the treatment and distinguish it from other factors with which it may be correlated. Specifically, we must distinguish between $d_i$, the treatment, and $z_i$, a variable that indicates which observations have been allocated to treatment or control. We seek to estimate the effect of $d_i$, and we assume that the treatment assignment $z_i$ has no effect on outcomes except insofar as it affects the value of $d_i$.

The term *exclusion restriction* or *excludability* refers to the assumption that $z_i$ can be omitted from the schedule of potential outcomes for $Y_i(1)$ and $Y_i(0)$. Formally, this

assumption may be written as follows. Let $Y_i(z, d)$ be the potential outcome when $z_i = z$ and $d_i = d$, for $z \in (0, 1)$ and for $d \in (0, 1)$. For example, if $z_i = 1$ and $d_i = 1$, the subject is assigned to the treatment group and receives the treatment. We can also envision other combinations. For example, if $z_i = 1$ and $d_i = 0$, the subject is assigned to the treatment group but for some reason does not receive the treatment. The exclusion restriction assumption is that $Y_i(1, d) = Y_i(0, d)$. In other words, potential outcomes respond only to the input from $d_i$; the value of $z_i$ is irrelevant. Unfortunately, this assumption cannot be verified empirically because we never observe both $Y_i(1, d)$ and $Y_i(0, d)$ for the same subject.

The exclusion restriction breaks down when random assignment sets in motion causes of $Y_i$ other than the treatment $d_i$. Suppose the treatment in our running example were defined as whether or not a woman council head presides over deliberations about village priorities. Our ability to estimate the effect of this treatment would be jeopardized if nongovernmental aid organizations, sensing that newly elected women will prioritize clean water, were to redirect their efforts to promote water sanitation to male-led villages. If outside aid flows to male-led villages, obviating the need for male village council leaders to allocate their budgets to water sanitation, the apparent difference between water sanitation budgets in councils led by women and councils led by men will exaggerate the true effect of the treatment, as defined above.[11] Even if it were the case that women council leaders have no effect on their own villages' budgets, the behavior of the NGOs could generate different average budgets in male- and female-led villages.

Asymmetries in measurement represent another threat to the excludability assumption. Suppose, for example, that in our study of Indian villages, we were to dispatch one group of research assistants to measure budgets in the treatment group and a different group of assistants to measure budgets in the control group. Each group of assistants may apply a different standard when determining what expenditures are to be classified as contributing to water sanitation. Suppose the research assistants in the treatment group were to use a more generous accounting standard—they tend to exaggerate the amount of money that the village allocates to water sanitation. When we compare average budgets in the treatment and control groups, the estimated treatment effect will be a combination of the true effect of female village heads on budgets and accounting procedures that exaggerate the amount of money spent on water sanitation in those villages. Presumably, when we envisioned the experiment and what we might learn from it, we sought to estimate only the first of these two effects. We wanted to know the effect of female leadership on budgets using a consistent standard of accounting.

11 Whether an excludability violation occurs depends on how a treatment effect is defined. If one were to define the effect of electing a woman to include the compensatory behavior of NGOs, this assumption would no longer be violated.



To illustrate the consequences of measurement asymmetry, we may write out a simple model in which outcomes are measured with error. Under this scenario, the usual schedule of potential outcomes expands to reflect the fact that outcomes are influenced not only by $d_i$, but also by $z_i$, which determines which set of research assistants measure the outcome. Suppose that among untreated units we observe $Y_i(0)^* = Y_i(0) + e_{i0}$, where $e_{i0}$ is the error that is made when measuring the potential outcome if an observation is assigned to the control group. For treated units, let $Y_i(1)^* = Y_i(1) + e_{i1}$. What happens if we compare average outcomes among treated and untreated units? The expected value of the difference-in-means estimator from equation (2.14) is

$$E\left[\frac{\sum_1^m Y_i}{m} - \frac{\sum_{m+1}^N Y_i}{N - m}\right] = E[Y_i(1)^* | D_i = 1] - E[Y_i(0)^* | D_i = 0]$$

$$= E[Y_i(1) | D_i = 1] + E[e_{i1} | D_i = 1] - E[Y_i(0) | D_i = 0] - E[e_{i0} | D_i = 0]. \quad (2.16)$$

Comparing equation (2.16) to equation (2.14) reveals that the difference-in-means estimator is biased when the measurement errors in the treated and untreated groups have different expected values:

$$E[e_{i1} | D_i = 1] \neq E[e_{i0} | D_i = 0]. \quad (2.17)$$

In this book, when we speak of a "breakdown in symmetry," we have in mind procedures that may distort the expected difference between treatment and control outcomes.

What kinds of experimental procedures bolster the plausibility of the excludability assumption? The broad answer is anything that helps ensure uniform handling of treatment and control groups. One type of procedure is double-blindness—neither the subjects nor the researchers charged with measuring outcomes are aware of which treatments the subjects receive, so that they cannot consciously or unconsciously distort the results. Another procedure is parallelism in the administration of an experiment: the same questionnaires and survey interviewers should be used to assess outcomes in both treatment and control groups, and both groups' outcomes should be gathered at approximately the same time and under similar conditions. If outcomes for the control group are gathered in October, but outcomes in the treatment group are gathered in November, symmetry may be jeopardized.

The exclusion restriction cannot be evaluated unless the researcher has stated precisely what sort of treatment effect the experiment is intended to measure and designed the experiment accordingly. Depending on the researcher's objective, the control group may receive a special type of treatment so that the treatment vs. control comparison isolates a particular aspect of the treatment. A classic example of a research design that attempts to isolate a specific cause is a pharmaceutical trial in

which an experimental pill is administered to the treatment group while an identical sugar pill is administered to the control group. The aim of administering a pill to both groups is to isolate the pharmacological effects of the ingredients, holding constant the effect of merely taking some sort of pill. In the village council example, a researcher may wish to distinguish the effects of female leadership of local councils from the effects of merely appointing non-incumbents to the headship. In principle, one could compare districts with randomly assigned women heads to districts with randomly assigned term limits, a policy that has the effect of bringing non-incumbents into leadership roles. This approach to isolating causal mechanisms is revisited again in Chapter 10, where we discuss designs that attempt to differentiate the active ingredients in a multifaceted treatment.

Protecting the theoretical integrity of the treatment vs. control comparison is of paramount importance in experimental design. In the case of the village budget study, the aim is to estimate the budgetary consequences of having a randomly allocated female village head, not the consequences of using a different measurement standard to evaluate outcomes in treatment and control villages. The same argument goes for other aspects of research activity that might be correlated with treatment assignment. For example, if the aim is to measure the effect of female leadership on budgets per se, bias may be introduced if one sends a delegation of researchers to monitor village council deliberations in women-headed villages only. Now the observed treatment effect is a combination of the effect of female leadership and the effect of research observers. Whether one regards the presence of the research delegation as a distortion of measurement or an unintended pathway by which assignment to treatment affects the outcome, the formal structure of the problem remains the same. The expected outcome of the experiment no longer reveals the causal effect we set out to estimate.

The symmetry requirement does not rule out cross-cutting treatments. For example, one could imagine a version of India's reservation policy that randomly assigned some village council seats to women, others to people from lower castes, and still others to women from lower castes. When we discuss factorial designs in Chapter 9, we will stress what can be learned from deploying several treatments in combination with one another. The point of these more complex designs is to learn about combinations of treatments while still preserving symmetry: randomly assigning treatments both alone and in combination with one another allows the researcher to distinguish empirically between having a female village head and having a female village head who is also from a lower caste.

Finally, let's revisit the case in which other actors intervene in response to your treatment assignments. For example, suppose that in anticipation of greater spending on water sanitation, interest groups devote special attention to lobbying village councils headed by women. Or it may go the other way: interest groups focus greater efforts on villages headed by men because they believe that's where they will meet the most resistance from budget makers. Whether interest group interference violated



the assumption of excludability depends on how we define the treatment effect. Interest group activity presents no threat to the exclusion restriction if we define the effect of installing a female council head to include all of the indirect repercussions that it could have on interest group activity. If, however, we seek to estimate the specific effect of having female council heads without any interference by interest groups, our experimental design may be inadequate unless we can find a way to prevent interest groups from responding strategically. These kinds of scenarios again underscore the importance of clearly stating the experimental objectives so that researchers and readers can assess the plausibility of the exclusion restriction.

### 2.7.2 Non-Interference

For ease of presentation, the above discussion only briefly mentioned an assumption that plays an important role in the definition and estimation of causal effects. This assumption is sometimes dubbed the Stable Unit Treatment Value Assumption, or SUTVA, but we refer to it by a more accessible name, non-interference.[12] In the notation used above, expressions such as $Y_i(d)$ are written as though the value of the potential outcome for unit $i$ depends only upon whether or not the unit itself gets the treatment (whether $d$ equals one or zero). A more complete notation would express a more extensive schedule of potential outcomes depending on which treatments are administered to other units. For example, for Village 1 we could write down all of the potential outcomes if only Village 1 is treated, if only Village 2 is treated, if Villages 1 and 2 are treated, and so forth. This schedule of potential outcomes quickly gets out of hand. Suppose we listed all of the potential outcomes if exactly two of the seven villages are treated: there would now be 21 potential outcomes for each village. Clearly, if our study involves just seven villages, we have no hope of saying anything meaningful about this complex array of causal effects unless we make some simplifying assumptions.

The non-interference assumption cuts through this complexity by ignoring the potential outcomes that would arise if subject $i$ were affected by the treatment of other subjects. Formally, we reduce the schedule of potential outcomes $Y_i(\boldsymbol{d})$, where $\boldsymbol{d}$ describes all of the treatments administered to all subjects, to a much simpler schedule $Y_i(d)$, where $d$ refers to the treatment administered to subject $i$.[13] In the context of our example, non-interference implies that the sanitation budget in one village is unaffected by the gender of the council heads in other villages. Non-interference is an assumption common to both experimental and observational studies.

12 The term "stable" in SUTVA refers to the stipulation that the potential outcomes for a given village remain stable regardless of which other villages happen to be treated. The technical aspects of this term are discussed in Rubin 1980 and Rubin 1986.

13 Implicit in this formulation of potential outcomes is the assumption that potential outcomes are unaffected by the overall pattern of actual or assigned treatments. In other words, $Y_i(\boldsymbol{z}, \boldsymbol{d}) = Y_i(z, d)$.

Is non-interference realistic in this example? It is difficult to say without more detailed information about communication between villages and the degree to which their budget allocations are interdependent. If the collection of villages were dispersed geographically, it might be plausible to assume that the gender of the village head in one village has no consequences for outcomes in other villages. On the other hand, if villages were adjacent, the presence of a woman council head in one village might encourage women in other villages to express their policy demands more forcefully. Proximal villages might also have interdependent budgets; the more one village spends on water sanitation, the less the neighboring village needs to spend in order to maintain its own water quality.

The estimation problems that interference introduces are potentially quite complicated and unpredictable. Untreated villages that are affected by the treatments that nearby villages receive no longer constitute an untreated control group. If women council heads set an example of water sanitation spending that is then copied by neighboring villages headed by men, a comparison between average outcomes in treatment villages and (semi-treated) control villages will tend to understate the average treatment effect as defined in equation (2.3), which is usually understood to refer to the contrast between treated potential outcomes and completely untreated potential outcomes. On the other hand, if female council heads cause neighboring villages headed by men to free ride on water sanitation projects and allocate less of their budget to it, the apparent difference in average budget allocations will exaggerate the average treatment effect. Given the vagaries of estimation in the face of interference, researchers often try to design experiments in ways that minimize interference between units by spreading them out temporally or geographically. Another approach, discussed at length in Chapter 8, is to design experiments in ways that allow the researcher to detect spillover between units. Instead of treating interference as a nuisance, these more complex experimental designs aim to detect evidence of communication or strategic interaction among units.

## SUMMARY

This chapter has limited its purview to a class of randomized experiments in which treatments are deployed exactly as assigned and outcomes are observed for all of the assigned subjects. This class of studies is a natural starting point for discussing core assumptions and what they imply for research design. The chapters that follow will introduce further assumptions in order to handle the complications that arise due to noncompliance (Chapters 5 and 6) and attrition (Chapter 7).

We began by defining a causal effect as the difference between two potential outcomes, one in which a subject receives treatment and the other in which the subject does not receive treatment. The causal effect for any given subject is not directly observable. However, experiments provide unbiased estimates of the average treatment effect (ATE) among all subjects when certain assumptions are met. The three assumptions invoked in this chapter are random assignment, excludability, and non-interference.



1. Random assignment: Treatments are allocated such that all units have a known probability between 0 and 1 of being placed into the treatment group. Simple random assignment or complete random assignment implies that treatment assignments are statistically independent of the subjects' potential outcomes.

   This assumption is satisfied when all treatment assignments are determined by the same random procedure, such as the flip of a coin. Because random assignment may be compromised by those allocating treatments or assisting subjects, steps should be taken to minimize the role of discretion.
2. Excludability: Potential outcomes respond solely to receipt of the treatment, not to the random assignment of the treatment or any indirect by-products of random assignment. The treatment must be defined clearly so that one can assess whether subjects are exposed to the intended treatment or something else.

   This assumption is jeopardized when (i) different procedures are used to measure outcomes in the treatment and control groups and (ii) research activities, other treatments, or third-party interventions other than the treatment of interest differentially affect the treatment and control groups.
3. Non-interference: Potential outcomes for observation $i$ reflect only the treatment or control status of observation $i$ and not the treatment or control status of other observations. No matter which subjects the random assignment allocates to treatment or control, a given subject's potential outcomes remain the same.

   This assumption is jeopardized when (i) subjects are aware of the treatments that other subjects receive, (ii) treatments may be transmitted from treated to untreated subjects, or (iii) resources used to treat one set of subjects diminish resources that would otherwise be available to other subjects. See Chapter 10 for a more extensive list of examples.

Random assignment is different from the other two assumptions in that it refers to a procedure and the manner in which researchers carry it out. Excludability and non-interference, on the other hand, are substantive assumptions about the ways in which subjects respond to the allocation of treatments. When assessing excludability and non-interference in the context of a particular experiment, the first step is to carefully consider how the causal effect is defined. Do we seek to study the effect of electing women to village council positions or rather the effect of electing women from a pool of candidates that consists only of women? When defining the treatment effect of installing a female village council head, is the appropriate comparison a village with male leadership, or a male-led village with no neighboring female-led villages? Attending to these subtleties encourages a researcher to design more exacting experimental comparisons and to interpret the results with greater precision.

Attentiveness to these core assumptions also helps guide experimental investigation, urging researchers to explore the empirical consequences of different research designs. A series of experiments in a particular domain may be required before a researcher can gauge whether subjects seem to be affected by the random assignment over and above the treatment (a violation of excludability) or by the treatments administered to other units (interference).

## SUGGESTED READINGS

Holland (1986) and Rubin (2008) provide non-technical introductions to potential outcomes notation. Fisher (1935) and Cox (1958) are two classic books on experimental design and analysis; Dean and Voss (1999) and Kuehl (1999) offer more modern treatments. See Rosenbaum and Rubin (1984) on the distinctive statistical properties of randomly assigned treatments.

## EXERCISES: CHAPTER 2

1. Potential outcomes notation:
   (a) Explain the notation "$Y_i(0)$."
   (b) Explain the notation "$Y_i(0) \mid D_i = 1$" and contrast it with the notation "$Y_i(0) \mid d_i = 1$."
   (c) Contrast the meaning of "$Y_i(0)$" with the meaning of "$Y_i(0) \mid D_i = 0$."
   (d) Contrast the meaning of "$Y_i(0) \mid D_i = 1$" with the meaning of "$Y_i(0) \mid D_i = 0$."
   (e) Contrast the meaning of "$E[Y_i(0)]$" with the meaning of "$E[Y_i(0) \mid D_i = 1]$."
   (f) Explain why the "selection bias" term in equation (2.15), $E[Y_i(0) \mid D_i = 1] - E[Y_i(0) \mid D_i = 0]$, is zero when $D_i$ is randomly assigned.
2. Use the values depicted in Table 2.1 to illustrate that $E[Y_i(0)] - E[Y_i(1)] = E[Y_i(0) - Y_i(1)]$.
3. Use the values depicted in Table 2.1 to complete the table below.
   (a) Fill in the number of observations in each of the nine cells.
   (b) Indicate the percentage of all subjects that fall into each of the nine cells. (These cells represent what is known as the joint frequency distribution of $Y_i(0)$ and $Y_i(1)$.)
   (c) At the bottom of the table, indicate the proportion of subjects falling into each category of $Y_i(1)$. (These cells represent what is known as the marginal distribution of $Y_i(1)$.)
   (d) At the right of the table, indicate the proportion of subjects falling into each category of $Y_i(0)$ (i.e., the marginal distribution of $Y_i(0)$).
   (e) Use the table to calculate the conditional expectation that $E[Y_i(0) \mid Y_i(1) > 15]$. (Hint: This expression refers to the expected value of $Y_i(0)$ given that $Y_i(1)$ is greater than 15.)
   (f) Use the table to calculate the conditional expectation that $E[Y_i(1) \mid Y_i(0) > 15]$.

| | $Y_i(1)$ | | | |
|---|---|---|---|---|
| $Y_i(0)$ | 15 | 20 | 30 | Marginal distribution of $Y_i(0)$ |
| 10 | | | | |
| 15 | | | | |
| 20 | | | | |
| Marginal distribution of $Y_i(1)$ | | | | 1.0 |



4. Suppose that the treatment indicator $d_i$ is either 1 (treated) or 0 (untreated). Define the average treatment effect among the treated, or ATT for short, as $\sum_1^N \tau_i d_i / \sum_1^N d_i$. Using the equations in this chapter, prove the following claim: "When treatments are allocated using complete random assignment, the ATT is, in expectation, equal to the ATE. In other words, taking expectations over all possible random assignments, $E[\tau_i \mid D_i = 1] = E[\tau_i]$, where $\tau_i$ is a randomly selected observation's treatment effect.
5. A researcher plans to ask six subjects to donate time to an adult literacy program. Each subject will be asked to donate either 30 or 60 minutes. The researcher is considering three methods for randomizing the treatment. One method is to flip a coin before talking to each person and to ask for a 30-minute donation if the coin comes up heads or a 60-minute donation if it comes up tails. The second method is to write "30" and "60" on three playing cards each, and then shuffle the six cards. The first subject would be assigned the number on the first card, the second subject would be assigned the number on the second card, and so on. A third method is to write each number on three different slips of paper, seal the six slips into envelopes, and shuffle the six envelopes before talking to the first subject. The first subject would be assigned the first envelope, the second subject would be assigned the second envelope, and so on.
   (a) Discuss the strengths and weaknesses of each approach.
   (b) In what ways would your answer to (a) change if the number of subjects were 600 instead of 6?
   (c) What is the expected value of $D_i$ (the assigned number of minutes) if the coin toss method is used? What is the expected value of $D_i$ if the sealed envelope method is used?
6. Many programs strive to help students prepare for college entrance exams, such as the SAT. In an effort to study the effectiveness of these preparatory programs, a researcher draws a random sample of students attending public high school in the United States, and compares the SAT scores of those who took a preparatory class to those who did not. Is this an experiment or an observational study? Why?
7. Suppose that an experiment were performed on the villages in Table 2.1, such that two villages are allocated to the treatment group and the other five villages to the control group. Suppose that an experimenter randomly selects Villages 3 and 7 from the set of seven villages and places them into the treatment group. Table 2.1 shows that these villages have unusually high potential outcomes.
   (a) Define the term *unbiased estimator*.
   (b) Does this allocation procedure produce upwardly biased estimates? Why or why not?
   (c) Suppose that instead of using random assignment, the researcher placed Villages 3 and 7 into the treatment group because the treatment could be administered inexpensively in those villages. Explain why this procedure is prone to bias.
8. Peisakhin and Pinto[14] report the results of an experiment in India designed to test the effectiveness of a policy called the Right to Information Act (RTIA), which allows citizens to inquire about the status of a pending request from government officials. In their study, the researchers hired confederates, slum dwellers who sought to obtain ration cards (which permit the purchase of food at low cost). Applicants for such cards must fill out a

14 Peisakhin and Pinto 2010.

form and have their residence and income verified by a government agent. Slum dwellers widely believe that the only way to obtain a ration card is to pay a bribe. The researchers instructed the confederates to apply for ration cards in one of four ways, specified by the researchers. The control group submitted an application form at a government office; the RTIA group submitted a form and followed it up with an official Right to Information request; the NGO group submitted a letter of support from a local nongovernmental organization (NGO) along with the application form; and finally, a bribe group submitted an application and paid a small fee to a person who is known to facilitate the processing of forms.

| | Bribe | RTIA | NGO | Control |
|---|---|---|---|---|
| Number of confederates in the study | 24 | 23 | 18 | 21 |
| Number of confederates who had residence verification | 24 | 23 | 18 | 20 |
| Median number of days to residence verification | 17 | 37 | 37 | 37 |
| Number of confederates who received a ration card within one year | 24 | 20 | 3 | 5 |

(a) Interpret the apparent effects of the treatments on the proportion of applicants who have their residence verified and the speed with which verification occurred.

(b) Interpret the apparent effects of the treatments on the proportion of applicants who actually received a ration card.

(c) What do these results seem to suggest about the effectiveness of the Right to Information Act as a way of helping slum dwellers obtain ration cards?

9. A researcher wants to know how winning large sums of money in a national lottery affects people's views about the estate tax. The researcher interviews a random sample of adults and compares the attitudes of those who report winning more than $10,000 in the lottery to those who claim to have won little or nothing. The researcher reasons that the lottery chooses winners at random, and therefore the amount that people report having won is random.

(a) Critically evaluate this assumption. (Hint: are the potential outcomes of those who report winning more than $10,000 identical, in expectation, to those who report winning little or nothing?)

(b) Suppose the researcher were to restrict the sample to people who had played the lottery at least once during the past year. Is it now safe to assume that the potential outcomes of those who report winning more than $10,000 are identical, in expectation, to those who report winning little or nothing?

10. Suppose researchers seek to assess the effect of receiving a free newspaper subscription on students' interest in politics. A list of student dorm rooms is drawn up and sorted randomly. Dorm rooms in the first half of the randomly sorted list receive a newspaper at their door each morning for two months; dorm rooms in the second half of the list do not receive a paper.

(a) University researchers are sometimes required to disclose to subjects that they are participating in an experiment. Suppose that prior to the experiment, researchers distributed a letter informing students in the treatment group that they would be receiving a newspaper as part of a study to see if newspapers make students more interested in politics. Explain (in words and using potential outcomes notation) how this disclosure may jeopardize the excludability assumption.

(b) Suppose that students in the treatment group carry their newspapers to the cafeteria where they may be read by others. Explain (in words and using potential outcomes notation) how this may jeopardize the non-interference assumption.

11. Several randomized experiments have assessed the effects of drivers' training classes on the likelihood that a student will be involved in a traffic accident or receive a ticket for a moving violation.[15] A complication arises because students who take drivers' training courses typically obtain their licenses faster than students who do not take a course.[16] (The reason is unknown but may reflect the fact that those who take the training are better prepared for the licensing examination.) If students in the control group on average start driving much later, the proportion of students who have an accident or receive a ticket could well turn out to be higher in the treatment group. Suppose a researcher were to compare the treatment and control group in terms of the number of accidents that occur within three years of obtaining a license.

(a) Does this measurement approach maintain symmetry between treatment and control groups?

(b) Would symmetry be maintained if the outcome measure were the number of accidents per mile of driving?

(c) Suppose researchers were to measure outcomes over a period of three years starting the moment at which students were randomly assigned to be trained or not. Would this measurement strategy maintain symmetry? Are there drawbacks to this approach?

12. A researcher studying 1,000 prison inmates noticed that prisoners who spend at least three hours per day reading are less likely to have violent encounters with prison staff. The researcher therefore recommends that all prisoners be required to spend at least three hours reading each day. Let $d_i$ be 0 when prisoners read less than three hours each day and 1 when prisoners read more than three hours each day. Let $Y_i(0)$ be each prisoner's potential number of violent encounters with prison staff when reading less than three hours per day, and let $Y_i(1)$ be each prisoner's potential number of violent encounters when reading more than three hours per day.

(a) In this study, nature has assigned a particular realization of $d_i$ to each subject. When assessing this study, why might one be hesitant to assume that $E[Y_i(0) \mid D_i = 0] = E[Y_i(0) \mid D_i = 1]$ and $E[Y_i(1) \mid D_i = 0] = E[Y_i(1) \mid D_i = 1]$?

(b) Suppose that researchers were to test this researcher's hypothesis by randomly assigning 10 prisoners to a treatment group. Prisoners in this group are required to go to the prison library and read in specially designated carrels for three hours each day for one week; the other prisoners, who make up the control group, go about their usual routines. Suppose, for the sake of argument, that all prisoners in the treatment group in fact read for three hours each day and that none of the prisoners

15 See Roberts and Kwan 2001.

16 Vernick et al. 1999.

in the control group read at all during the week of the study. Critically evaluate the excludability assumption as it applies to this experiment.

(c) State the assumption of non-interference as it applies to this experiment.

(d) Suppose that the results of this experiment were to indicate that the reading treatment sharply reduces violent confrontations with prison staff. How does the non-interference assumption come into play if the aim is to evaluate the effects of a policy whereby all prisoners are required to read for three hours?

# CHAPTER 3

# Sampling Distributions, Statistical Inference, and Hypothesis Testing

Rigorous quantification of uncertainty is a hallmark of scientific inquiry. When analyzing experimental data, the aim is not only to generate unbiased estimates of the average treatment effect but also to draw inferences about the uncertainty surrounding these estimates. Among the most attractive features of experimentation is that random allocation of treatments is a reproducible procedure. Reproducibility allows us to assess the sampling distribution, or collection of estimated ATEs that could have come about under different random assignments in order to better understand the uncertainty associated with the experiment we conducted. One objective of this chapter is to explain how experimental design affects the sampling distribution. We consider ways of designing experiments so as to reduce sampling variability, and we call attention to the fact that the sampling distribution may change markedly depending on the procedures used to randomly allocate subjects to treatment and control conditions.

A second objective is to guide the reader through the calculation and interpretation of key statistical results. When analyzing an experiment, you should consider both the estimated ATE and the uncertainty with which it is estimated. Unless you have prior information about the value of the ATE, the experimental estimate is one's best guess of the true treatment effect, but this guess may be close to or far from the true average causal effect. Statisticians commonly assess uncertainty in two ways. One method is to investigate whether the experimental results are sufficiently informative to refute a determined skeptic who insists that there is no treatment effect whatsoever. Another approach is to identify a range of values that probably bracket the true average treatment effect. This chapter introduces a flexible set of statistical techniques that may be used to assess uncertainty across a wide array of different experimental designs.