Observation & Experiment

*An Introduction to Causal Inference*

PAUL R. ROSENBAUM

Observation and Experiment

Observation and Experiment

An Introduction to Causal Inference

PAUL R. ROSENBAUM

Cover illustrations all Thinkstock by Getty Images: magnifying glass icon by Panptys; pie chart icon by macrovector; test tube icon by anttohoho.

Cover design by Lisa Roberts

CONTENTS

PREFACE

It is often said, "You cannot prove causality with statistics." One of my professors, Frederick Mosteller, liked to counter, "You can only prove causality with statistics." He had a particular statistical method in mind when saying this, the randomized experiments and randomized clinical trials that are described in Part I. In a randomized trial, a coin is flipped to assign each person to treatment or control. In a precise sense, randomized experiments warrant inference about causal effects on populations of individuals, where it is not possible to know the effect on any one individual. Randomized experimentation was invented by Sir Ronald Fisher in the 1920s and 1930s.[1]

Randomized trials are not always practical, not always ethical. You cannot randomize what you do not control, and many interesting treatments are outside the investigator's control. It is important to know the effects of fiscal policy on unemployment and economic growth, but in a democratic society, fiscal policy is not under the control of an investigator. In a civilized society, you can treat a person only with their informed consent, and even with consent you can treat a person only if the treatment is either harmless or holds a realistic prospect of benefit for the person treated. You cannot force someone to smoke cigarettes to discover whether smoking causes lung cancer. Many interesting treatments are outside experimental control, harmful, or both. Given that you cannot randomize, how can you study

the effects of such treatments? Part II discusses observational studies—that is, nonrandomized studies—of the effects caused by treatments.

My goal in this book is to present the concepts of causal inference clearly, with reasonable precision, but with a minimum of technical material. Difficult concepts are made accessible without dilution or distortion by focusing on the simplest nontrivial case, by omitting detail and generalization that are not essential, by presenting needed background material, and by introducing notation at a measured pace. Concepts are introduced through examples that are first discussed in English, then these same concepts are restated in precise terms. In a slogan, the goal is accessible precision.

Although I have tried to minimize the technical material, the basic concepts of causal inference do require a little notation, a little probability, and a little statistics—but really only a little. Because only a little of each is needed, I have tried to make the book self-contained by including what is needed. The endnotes contain pointers to technical material, for the people who like technical material, but there is no need to look at the endnotes. My hope and intention is that the book will be comfortable to read regardless of the reader's background. In this book, the mathematics never rises much above coin flips. Everyone can do coin flips.[2]

The title, "Observation and Experiment," has a history. It was the title of book 3, chapter 7, of John Stuart Mill's book *A System of Logic* (1843), a bible of nineteenth-century British empiricism.[3] To discover the effects of treatments, Mill advocated comparing identical people under alternative treatments. Observational studies that compare identical twins under alternative treatments owe something to Mill's *Logic*.[4]

In 1935, while introducing randomized experimentation, Fisher attacked Mill's notion, describing it as "a totally impossible requirement," essentially because identical people do not exist.[5] Fisher's notion, random assignment of treatments, would not make people the same; rather, it would make the treatment people received unrelated to everything that makes people different. A triumph of Fisher's method occurred in the 1954 field trial of Salk's poliomyelitis vaccine, in which more than 400,000 children in the United States were randomly assigned to either vaccine or placebo, an instructive story to which we will return.[6]

Sir Austin Bradford Hill was an early advocate of randomized trials of human subjects, but he was also, with Sir Richard Doll, the author in 1954 of an early, influential observational study claiming that smoking causes

cancer of the lung.[7] Hill published a paper entitled "Observation and Experiment" in 1953.[8] In 1964, the U.S. Public Health Service published *Smoking and Health: Report of the Advisory Committee to the Surgeon General,* which concluded, largely based on observational studies of human populations, that cigarette smoking causes lung cancer.[9] A statistician serving as a member of this advisory committee, William G. Cochran, published in 1965 a survey of the statistical principles guiding observational studies, entitled "The Planning of Observational Studies of Human Populations."[10] The title, "Observation and Experiment," marks the modern distinction between randomized experiments and observational studies.

Most parts of my story are true. However, in an effort to make notation human and humane, Harry and Sally appear now and then. They are fictions.

Academic writing tends to be heavy. Statistical writing tends to be heavy. Academic statistical writing tends to be very heavy. One can sympathize with Italo Calvino, who said that he devoted his life to the subtraction of weight: "above all I have tried to remove weight from the structure of stories and from language." Calvino wrote, "Lightness for me goes with precision and determination, not with vagueness and the haphazard. Paul Valéry said: 'One should be light like a bird, not light like a feather.' "[11]

In the interests of subtracting weight, I hope you will forgive me if, in our chat, I speak now and then of the two of us, and, of course, poor Harry.

READING OPTIONS

In writing, I have drawn two red lines through causal inference, dividing the subject into three parts. To the left of the first red line are topics that can be illustrated with a scientific example, chatted about in English, illuminated with high school algebra or coin flips, or illustrated with a simulated example. These topics appear in the text of this book. They are not necessarily elementary topics; rather, they are topics that can be discussed in a certain way.

Between the two red lines are topics that require familiarity and comfort with the concept of conditional probability for discrete random variables—that is, for dice, coin flips, and card games. The endnotes to the book often restate and expand upon topics using conditional probability. A person who has taken an undergraduate course in probability is likely to have encountered conditional probability a few weeks into the semester. If you know conditional probability, then look at the endnotes—they might provide added insight. If you do not know conditional probability, then do not look at the endnotes—they won't provide added insight. In addition, the text oversimplifies a few minor topics, correcting the oversimplifications in the endnotes.[1]

Beyond the second red line are topics that are a little too technical to discuss in this book. The endnotes provide pointers to places where you can read more if you are so inclined. There is also a list of suggested reading at the back of the book.

Sections with an asterisk (*) explain in English some ideas that require additional technical detail to fully explain. A section with an asterisk is not difficult, just incomplete. Such a section may offer the gist of an idea, but if you want more than that, you will have to dig into the references. You may view these sections as sketches of a foreign landscape or as invitations to travel, at your discretion. You may skip these sections without encountering problems.

I say most things several times: at first in English, and later with additional devices. If one of these variations makes sense and another is puzzling, then it is safe to continue reading. Truly, when has it ever been unsafe to continue reading? If you encounter an insurmountable obstacle in your reading, walk around it, continue to the next section or chapter with no concern or regret, and never look back. In all likelihood, you have walked around the illusion of an insurmountable obstacle. Perhaps the only insurmountable obstacle is the illusion of an insurmountable obstacle.[2]

I cover most topics from the beginning. The reader is not assumed to be familiar with basic statistical concepts, such as hypothesis testing and confidence intervals; rather, these concepts are explained within the context of randomized experiments in Chapter 3. I introduce notation slowly, with some repetition. If statistical concepts are entirely familiar to you and if you pick up notation easily, then I would encourage you to move through Chapters 2 and 3 fairly quickly.

For an introduction to causal inference, read Chapters 1 through 9, skipping sections with an asterisk and ignoring endnotes.

If you want to quickly reach a particular topic, pick up the notation and terminology in Chapters 2 and 5, and then jump to your topic. Chapter 3 depends strongly upon Chapter 2, and Chapters 6 through 13 depend somewhat upon Chapter 5. Chapter 10 depends upon Chapter 9, and Chapters 11 and 13 make occasional references to Chapter 9. Aside from this, the connections among chapters are limited. A later chapter may refer to a scientific example from an earlier chapter, but at that moment you can flip back to the earlier chapter.

Sections with an asterisk and endnotes provide a gentle introduction and invitation to ideas in the technical literature. The endnotes also contain references to statistical software packages in R and bits of code. You can obtain R and its packages for free at https://cran.r-project.org/. These packages often contain data from scientific studies used to illustrate methodology in

articles about statistical methodology. If you follow the trail of bread crumbs in the endnotes, you can try things out, reproducing analyses from technical papers.

Who are Harry and Sally? I have taught statistics for many decades and have often had the following experience. I say to a student, "If person i received treatment, then we would see the response of person i to treatment, but not the response of person i to control, so we cannot see the causal effect for person i."[3] The student stares at me in horror like I just landed from Saturn breathing fire. Then I say, "If Harry received aggressive treatment at the emergency room, we would see Harry's response to aggressive treatment, but we would not see his response to less aggressive treatment, so we cannot see if he benefited from aggressive treatment." The student responds, "Got it." Then I say, "But it isn't really about Harry or aggressive treatment or the emergency room; it's general." The student glares at me and says, "Yes, of course." I'm thinking, "If this student can abstract the general argument from a single sentence about Harry, then what was so hard about person i?" The student is thinking, "If that's all he meant, then why didn't he just say so the first time?" So, in this book, I just say so the first time. The typical person has an astonishing capacity for rigorous abstract thought but does not realize it. But don't worry—I'm not going to mention this.

LIST OF EXAMPLES

- Randomized trial of emergency treatment for septic shock (Chapters 1–3)
- Can a preference be irrational? (Chapter 4)
- Polio and the Salk vaccine (Chapter 4)
- Why do babies cry? (Chapter 5)
- Delirium in the hospital (Chapter 5)
- Bereavement and depression (Chapter 6)
- Prenatal malnourishment and cognitive development: the Dutch famine (Chapter 6)
- Diethylstilbestrol and vaginal cancer (Chapter 6)
- The 2010 Chilean earthquake and post-traumatic stress (Chapters 6 and 9)
- Father's occupation and lead in his children's blood (Chapters 7 and 9)
- Restrictions on handgun purchases and violent crime (Chapter 7)
- Fetal alcohol syndrome (Chapter 7)
- Antibiotics and death from cardiovascular causes (Chapter 8)

If a scientific man be asked what is truth, he will reply—if he frame his reply in terms of his practice and not of some convention—that which is accepted upon adequate evidence. And if he is asked for a description of adequacy of evidence, he certainly will refer to matters of observation and experiment . . . To exclude consideration of these processes [of inquiry] is thus to throw away the key to understanding knowledge and its objects.

—JOHN DEWEY, *Essays in Experimental Logic*[1]

The world shows up for us. But it does not show up for free . . .
We achieve access to the world around us through skillful engagement; we acquire and deploy the skills needed to bring the world into focus.

—ALVA NOË, *Varieties of Presence*[2]

What does it mean for experiments, if they can be beautiful? And what does it mean for beauty, if experiments can possess it?

—ROBERT P. CREASE, *The Prism and the Pendulum*[3]

Part I

RANDOMIZED EXPERIMENTS

A Randomized Trial

Emergency Treatment of Septic Shock

Septic shock occurs when a widespread infection leads to very low blood pressure. It is often lethal, particularly in young children and the elderly. At first, the infection produces chills, weakness, rapid heartbeat, and rapid breathing. Then it damages small blood vessels, so they leak fluid into nearby tissues. The heart has increased difficulty pumping, blood pressure drops, and less blood reaches vital organs. Each year in the United States, there are more than 750,000 cases of septic shock. The initial treatment for septic shock typically occurs in a hospital emergency room, and 20% of patients may die.

What is the best way to treat septic shock? Rivers and colleagues proposed an aggressive, six-hour protocol for treatment of septic shock.[1] In 2014, the *New England Journal of Medicine* reported the results of a randomized clinical trial comparing the aggressive protocol with a less-aggressive, six-hour protocol of standard therapy.[2] Where the aggressive protocol required or prompted certain actions—the placement of a catheter in a large vein to monitor blood pressure and oxygen, and packed red-cell transfusions—the less-aggressive protocol permitted but was less quick to prompt these activities.

For brevity, these two treatments will be called aggressive treatment and less-aggressive treatment. The trial was conducted by the Protocolized Care for Early Septic Shock (ProCESS) Investigators, a collaboration among researchers working in 31 emergency departments in the United States.

In a randomized trial, a fair coin is flipped to decide the treatment for the next patient.[3] For the ProCESS Trial, the coin flips assigned 439 patients to the aggressive protocol and 446 others to the less-aggressive protocol. Had you and I been patients in this clinical trial, the two coin flips for you and me might have given you the aggressive treatment and me the less-aggressive treatment; however, they might instead, with the same probability, have assigned me to aggressive treatment and you to the less-aggressive treatment. You and I are different, but the treatment assignments have a certain kind of equity, a certain of kind of symmetry, that is quite apart from the difference between you and me.

Causal Questions

What Are Causal Questions?

What is the best way to treat septic shock? This is a causal question, a question about the effects caused by treatments. The question begins with two possible treatments for septic shock.[4] For a specific patient—you or I, say—it asks, What would happen to this patient under the first treatment? What would happen to the patient under the second treatment? Would the patient fare better under the first rather than the second treatment? Would the outcome be the same under the two treatments? Causal effects are comparisons of potential outcomes under alternative treatments.[5]

Why is causal inference difficult? Why is the random assignment of treatments helpful in causal inference?

Causal inference is difficult because each patient receives one treatment, not both treatments. In the ProCESS Trial, some patients who arrived at the emergency department received the aggressive protocol, and others received the less-aggressive protocol. So for any one patient we see how that patient fared under the one treatment that patient received. We cannot see how that patient would have fared under the alternative treatment the patient did not receive. Hence, we can never see, for any patient, the comparison

of how this one patient would have fared under the two alternative treatments. We never see the causal effect because the causal effect is the comparison of these two potential outcomes the patient would have exhibited under the two alternative treatments. Causal inference is difficult because it is about something we can never see.

We cannot see the effect of aggressive versus less-aggressive treatment for any one patient. Can we see something else? In the ProCESS Trial, there was a finite population made up of the $885 = 439 + 446$ patients who received either the aggressive treatment or the less-aggressive treatment. Think of the 885 patients in this finite population as analogous to a population of 885 people living in a small town, where the population is divided roughly in half at random by repeated flips of a fair coin. Each patient in this population has a potential outcome under aggressive treatment and a potential outcome under less-aggressive treatment, but we see one or the other, never both. The situation is different for the population of 885 patients. The 439 patients who received the aggressive treatment are a random sample—roughly half of the population of 885 patients. The 446 patients who received the less-aggressive treatment also are a random sample—also roughly half of the same population of 885 patients. If you see outcomes for a random half of a population, and if the population is not very small, then you can say quite a bit about outcomes for the whole population. It is partly that you saw half the population, but more importantly you saw a random half—the half you did not see could not be that different from the half you saw. From our two random halves of the population, halves that received different treatments, we will be able to say quite a bit about how all 885 patients would have fared if they all had received aggressive treatment or if they all had received less-aggressive treatment. That is, we will be able to say quite a bit about the effect of the treatment on the 885 people as a group, even though we are limited in what we can say about any one of them. Chapter 3 will make this intuition precise.

Causal Questions Distinguish Covariates and Outcomes

A covariate is a quantity determined prior to treatment assignment.[6] In the ProCESS Trial, the age of the patient at the time of admission to the emergency room was a covariate. The gender of the patient was a covariate.

Whether the patient was admitted from a nursing home was a covariate. A very important covariate was the Acute Physiology and Chronic Health Evaluation (APACHE) II score at the time of admission, a numerical summary of the severity of the patient's acute health problems.[7] An APACHE II score is a snapshot, a moment in time, a still photo of a running horse; it describes the patient's acute problems right now, and a different APACHE II score describes this same patient's acute problems an hour or a day or a week later. Higher APACHE II scores indicate more severe acute health problems.

Notice that a variable is stamped with a date: the APACHE II score at the time of admission to the emergency room is a different variable than the APACHE II score a week later. An APACHE II score measured before treatment assignment is a covariate. An APACHE II score measured a week after treatment is not a covariate but an outcome. Two variables can have a similar sound when described in English—the APACHE II score an hour before treatment, and the APACHE II score a day after treatment—yet despite the similar sound, they may have different structures, one a covariate and the other an outcome. In distinguishing covariates and outcomes, time is important—indeed, the role that time plays in distinguishing covariates and outcomes is central to the role time plays in cause and effect.

Because a covariate is determined before the treatment assignment, the covariate does not change when the patient is assigned to one treatment or another. Assigning a patient to the aggressive protocol rather than the less-aggressive protocol will not change the patient's age upon admission, the patient's gender, or whether the patient came to the emergency room from a nursing home. In exactly the same way, though now with greater potential for misunderstanding, assigning a patient to the aggressive protocol rather than the less-aggressive protocol will not change the patient's APACHE II score upon admission to the emergency room before treatment assignment. Of course, if the aggressive treatment were more effective than the less-aggressive treatment, then the assignment to the aggressive treatment might affect an APACHE II score recorded three hours after the start of treatment—that is a different variable, recorded at a different time, and it is an outcome not a covariate. Assigning a treatment now can alter the future, but it cannot alter the past; it cannot alter a covariate.

An outcome is a quantity determined after treatment assignment, hence a quantity that may have been affected by receiving one treatment rather than

the other. An outcome exists in two versions: its value if the patient receives the aggressive protocol, and its value if the patient receives the less-aggressive protocol. The primary outcome in the ProCESS Trial was in-hospital mortality by 60 days after treatment assignment. In terms of this primary outcome, success meant being discharged alive from the hospital before 60 days or being alive in the hospital at 60 days, and failure meant dying in the hospital before 60 days.[8]

That outcome might be different for a particular patient, say you or me, depending on whether the patient received the aggressive protocol or the less-aggressive protocol. If aggressive treatment were of great benefit, then perhaps this particular patient might have been discharged alive before 60 days if given aggressive treatment, but might have died in the hospital before 60 days if given less-aggressive treatment. That is to say, aggressive treatment might cause the patient to be discharged alive, whereas less-aggressive treatment might cause the patient to die. Or perhaps aggressive treatment is of no benefit, meaning it does not alter survival. If aggressive treatment were of no benefit to anyone, then, by the definition of "no benefit," changing the treatment assigned to any of these patients would not have changed whether that patient lived or died. Causal effects are comparisons of the potential outcomes patients would exhibit under alternative treatments.[9]

Covariates That Were Not Measured; Potential Outcomes That Were Not Seen

If we accurately measure a covariate before the treatment assignment, then we observe the one and only value that this covariate has. If we forget or are unable to measure a covariate, then it still has only one value, but we do not know that value; such a covariate is said to be "not observed," or more commonly but less gracefully, "unobserved." At the end of the day, scientific arguments about what causes what are almost invariably arguments about some covariate that was not measured or could not be measured. Someone claims that differing outcomes in treated and control groups stem not from a treatment effect but from comparing people who were not comparable before treatment. John Stuart Mill wanted to compare identical people under alternative treatments—that is, he wanted there to be no unmeasured pretreatment

differences between people receiving treatment or control; however, Mill did not provide us with a method to create this ideal situation.[10] As we will see, a randomized treatment assignment provides a strong promise about covariates that were not measured.

If we accurately measure an outcome, we see one of its two potential values: the value that occurs under the treatment the patient actually received. We can never see the outcome a patient would have exhibited under the treatment the patient did not receive. We might see that a patient who received aggressive care was discharged alive before 60 days, but we cannot see whether this patient would have been discharged alive before 60 days had the patient received less-aggressive care. True, someone else might have received less-aggressive care and have died in the hospital, but someone else is someone else. Perhaps that other person was much sicker upon admission to the emergency room in some way that was not measured—sicker in terms of an unobserved covariate—and the less-aggressive care had nothing to do with the patient's death.

The distinction between covariates and outcomes is basic in causal inference. Failure to distinguish covariates and outcomes quickly leads to confusion, errors, and false conclusions.[11] There are many genuine, sometimes unavoidable difficulties in causal inference; however, it is easy to distinguish covariates and outcomes, so it is easy to avoid the difficulties that come from confusing them.

Perhaps the distinction between covariate and outcome is most vivid, most palpable, in Robert Frost's poem "The Road Not Taken" (1916):

> Two roads diverged in a yellow wood,
> And sorry I could not travel both
> And be one traveler, long I stood
> And looked down one as far as I could
> To where it bent in the undergrowth;

Frost creates the mood attending a decision, one whose full consequences we cannot see or anticipate: "Knowing how way leads on to way," we will not see the road not taken.

As it was for Frost in a yellow wood, so it is for a patient at risk of death in the ProCESS Trial, and so it will be in every causal question.

Randomization Balances Covariates

Absent Randomization, Treated and Control Groups May Differ before Treatment

The ProCESS Trial randomly assigned patients to either an aggressive protocol or a less-aggressive protocol; in effect, fair coins were flipped to assign treatments. Random assignment does not occur naturally. In all likelihood, treatment assignments that occur naturally produce a biased comparison, one in which better outcomes (such as fewer deaths) occur in one treatment group, but the better outcomes might not be caused by that treatment. It would be natural to treat less aggressively a patient whose illness is less severe, or to treat more aggressively a patient who might survive but whose illness is more severe, or perhaps to treat less aggressively a patient who has no hope of surviving. It would be natural to select patients in this way for aggressive treatment, even if it is not known whether aggressive treatment is beneficial or harmful or for whom it is beneficial.

If severely ill patients receive aggressive treatment and less severely ill patients receive less-aggressive treatment, then the aggressive treatment could look harmful when it is actually beneficial. Severely ill patients are the ones most likely to die. If severely ill patients are concentrated in the aggressive treatment group, with the less severely ill patients concentrated in the less-aggressive treatment group, then the aggressive treatment group would be expected to have more deaths, even if both groups receive exactly the same treatment. If severely ill patients are concentrated in the aggressive treatment group, then in its contest with less-aggressive treatment the aggressive treatment given to the sickest patients starts at a disadvantage. It starts the race many paces behind: even if it ran faster—saved more lives—it still might come in second with a higher death rate because its patients were sicker to begin with. If you want to know whether the aggressive treatment saves more lives than the less-aggressive treatment, then you need an equitable comparison. Equitable comparisons are rare in nature.

Randomization seems equitable, as the flip of a fair coin favors neither treatment. But the concepts of equity and fairness are out of place here: we have reason to be fair to people, not to treatments. Thus, although fairness is a familiar concept, the one we need is slightly different. Our concepts of

fairness and equity often involve concepts of symmetry. In many contexts, an arrangement is equitable if it is symmetrical, the same for everyone. It is the symmetry of random assignment that does the work in experiments: there would be no reason to be "fair" to treatments if we knew what effects they cause. Is there something that we can see that exhibits this symmetry?

Randomization Balances Observed Covariates

In the ProCESS Trial, random assignment of treatments meant that each patient was equally likely to receive aggressive treatment or less-aggressive treatment. A 72-year-old female patient with an APACHE II score of 30 was equally likely to receive aggressive treatment or less-aggressive treatment; therefore, aside from chance—aside from an unlucky sequence of coin flips in assigning treatments—72-year-old female patients with APACHE II scores of 30 are expected to occur with equal frequencies in both treatment groups. The key here is that the chance of receiving aggressive treatment does not change with the attributes that described the patients before treatment. The key is that fair coins ignore the attributes of the patients when assigning treatments to them.

Table 1.1 compares the aggressive and less-aggressive treatment groups before treatment assignment. That is, it compares the groups in terms of several measured covariates. For instance, the mean or average age of the 439 patients in the aggressive treatment group was 60 years, whereas the mean age of 446 patients in the less-aggressive treatment group was 61 years—not much of a difference there. Patients developed septic shock from various underlying infections (for instance, pneumonia), but the two treatment groups have similar frequencies of these underlying infections. As expected when studying septic shock, blood pressure was low, and serum lactate was high; however, both of these covariates looked similar in the aggressive and less-aggressive treatment groups. In terms of these nine covariates, the two groups look fairly comparable; that is, the nine covariates look balanced. Given that the two groups looked similar before treatment, it seems reasonable to compare the outcomes of aggressive and less-aggressive treatment.

One remarkable thing about Table 1.1 is that the groups look fairly comparable in terms of these nine covariates. A second, more important, more remarkable thing about the table is that the nine covariates were not used in

Table 1.1. Covariate balance in the ProCESS Trial

Treatment group	Aggressive	Less aggressive
Sample size (number of patients)	439	446
Age (years, mean)	60	61
Male (%)	53	56
Came from nursing home (%)	15	16
Sepsis source		
Pneumonia (%)	32	34
Urinary tract infection (%)	23	20
Intra-abdominal infection (%)	16	13
APACHE II score (mean)	20.8	20.6
Systolic blood pressure (mean)	100	102
Serum lactate (mmol / liter, mean)	4.8	5

APACHE II = Acute Physiology and Chronic Health Evaluation.

creating the two treatment groups: the balance in Table 1.1 is just luck. Yet the ProCESS Investigators anticipated that luck would balance these covariates. Take a chance once, and you may be surprised by what happens. Take the same chance independently 885 times, and you may see a desired property—here, covariate balance—that was very likely to happen. Luck can be very reliable.[12]

Randomization Balances Unobserved Covariates

Table 1.1 shows the two groups were fairly comparable, adequately comparable, in terms of the nine measured covariates before the start of treatment. If your sole goal was to compare two groups comparable in terms of these nine covariates, then you could assign treatments to exhibit greater comparability than seen in Table 1.1. For instance, you could make the percentage of males very close to 54.5%, rather than 53% and 56%. To tidy up the small differences, you would have to use the nine covariates in assigning treatments, forcing balance. In some contexts, you might be able to tidy up a table like Table 1.1 while still assigning treatments at random.[13] However, if you tidied up Table 1.1 for nine covariates by not randomly assigning treatments, then you would have taken a big step backward, not a small step forward.

The covariate balance in Table 1.1 was achieved by coin flips that made no use of the nine covariates. Table 1.1 was produced by a fair lottery. The covariates were measured, so we can see that randomization balanced them by calculating their means, but random assignment ignores the covariates in favor of flipping coins. So if age had not been measured, it would not have mattered: the same coin flips that actually occurred in the ProCESS Trial would have produced the same treatment assignments and the same mean ages, 60 and 61, seen in Table 1.1. In other words, we have reason to believe that some other covariate, a covariate that was not measured by the ProCESS Investigators, was balanced in the same way that the covariates in the table are balanced. This claim about balancing unmeasured covariates is a theoretical claim, a claim based on probability theory. Yet Table 1.1 provides evidence that the theoretical claim has traction in reality.

It is common, indeed typical, to criticize nonrandomized or observational studies of treatment effects by claiming that the two groups being compared were not comparable before treatment. Although the groups may appear to be comparable in terms of a few measured covariates, the critics argue that appearances deceive: the groups look comparable, but they differ in terms of an unmeasured covariate. The differing outcomes are not effects of the treatment, the critics continue, but rather are consequences of comparing people who were not comparable at the outset. Criticism of this form is sometimes reasonable, sometimes not, as we will see in Part II. More often, criticism of this form can be difficult to evaluate as either reasonable or unreasonable.

Typically, the special tools of sensitivity analysis in Chapter 9 are needed to clarify what the critic is saying and whether it is plausible in light of the available data. In designing observational studies, investigators attempt to preempt such criticism, as discussed in Chapters 6, 8, and 10. Because the ProCESS Investigators assigned treatments at random, they have a strong response to such a critic: randomization is expected to balance such unobserved covariates just as it balanced the observed covariates seen in Table 1.1. This is because it achieves covariate balance by holding a fair lottery, not by tinkering with specific observed covariates. In other words, because they randomized their patients, the ProCESS Investigators could respond to such a critic with "You are mistaken." The precise sense in which this is true is the topic of Chapter 3.

Is Aggressive Treatment Better?

Is the aggressive protocol of treatment for septic shock better than the less-aggressive protocol? Table 1.2 displays the primary outcome, in-hospital death by 60 days. The in-hospital mortality rate was 21.0% with the aggressive protocol, and 18.2% with the less-aggressive protocol. On the surface, the aggressive protocol appears no better and perhaps somewhat worse than the less-aggressive protocol. Some care is needed in interpreting these two mortality rates because they refer to different people. The 21.0% mortality rate refers to the 439 people randomly picked for aggressive treatment, and the 18.2% mortality rate refers to the 446 people randomly picked for less-aggressive treatment. The groups received different treatments, but they also contain different people.

We saw in Table 1.1 that these groups of patients are similar as groups but are not identical; of course they could differ in ways not recorded in Table 1.1. Perhaps the difference between the two mortality rates is due to chance, the turn of a coin that assigned one patient to aggressive treatment and the next patient to less-aggressive treatment. Is the difference between 21.0% and 18.2% too large to be due to chance, that is, too large to be due to an unlucky sequence of coin flips in assigning treatments? If the aggressive treatment conferred no benefit but did no harm in comparison with the less-aggressive treatment, could the coin flips that assigned treatments to patients produce mortality rates as different as 21.0% and 18.2%? Or can we be confident that less-aggressive treatment saves lives in comparison with aggressive treatment? Chapter 3 will discuss how these questions are answered.

We say that there is no difference in the effects of two treatments if every patient would fare the same under the first and second treatments. Let us say that more precisely. In the ProCESS Trial, "no difference in effect" means

Table 1.2. In-hospital mortality outcomes in the ProCESS Trial

| Treatment group | *In-hospital 60-day mortality* | | | |
	In-hospital death	*Other*	*Total*	*Death rate (%)*
Aggressive	92	347	439	21.0
Less aggressive	81	365	446	18.2
Total	173	712	885	

that every patient who would die in the hospital by 60 days with aggressive treatment would also die in the hospital by 60 days with less-aggressive treatment; conversely, every patient who would die in the hospital by 60 days with less-aggressive treatment would also die in the hospital by 60 days with aggressive treatment. If this were true, then the two potential outcomes would be equal for each patient. If this were true in Frost's poem, the two paths would always end in the same place. This is the null hypothesis of no difference in the effects of the two treatments being compared, or, more precisely, Fisher's sharp null hypothesis of no difference in effects of the two treatments.[14]

The precise meaning of Fisher's sharp null hypothesis will be developed in Chapter 3, but in essence it means that each and every one of the 885 patients experienced no effect: some lived, others died, but no one lived or died because they received aggressive rather than less-aggressive treatment. In the interest of concise expression, we often refer to this as the null hypothesis of no effect, but that is just a shorthand (a sometimes confusing or misleading shorthand) for no difference in the effects of the two treatments being compared. As we will see in Chapter 3, if there were no difference in the effects of aggressive and less-aggressive protocols, and if the coins were flipped again and again, one lottery after another, repeatedly splitting the 885 patients into groups of size 439 and 446, then about one-third of these lotteries would produce splits with a difference in mortality rates as large or larger than the observed 21.0% and 18.2%.[15] In that sense, the difference we saw could be due to chance—the coin flips alone—in the absence of any difference in the effects of the two treatment protocols. Certainly, the ProCESS Trial gives no reason to believe the aggressive protocol is better than the less-aggressive protocol in terms of in-hospital mortality, but neither does it give strong evidence that the aggressive protocol is worse. Again, all of this is developed carefully in Chapter 3.

The 2 × 2 Table and Fisher's Hypothesis

Before we leave Table 1.1, let us consider what Fisher's hypothesis of no effect would mean for the table. The hypothesis says that changing the treatment a patient receives would not change that patient's in-hospital mortality outcome. In terms of Table 1.1, the hypothesis means that changing a pa-

tient's treatment, or row of the table, would not change the patient's outcome or column of the table. If Fisher's hypothesis of no difference in effects were true, then the 173 patients in the first column would be in the first column whether they were assigned to aggressive treatment or to less-aggressive treatment. In parallel, if Fisher's hypothesis were true, then 712 patients in the second column would remain in the second column even if their treatments were switched. If changing the treatment a patient receives changes the column this patient is in, then one treatment would save this patient who would die under the alternative treatment, so Fisher's hypothesis of no difference in effects is false. The imagined repeated lotteries in the previous paragraph would pick 439 of the 885 patients at random for the first row, but if Fisher's hypothesis were true, it would not matter which 439 patients are picked so far as the total in the first column is concerned: every lottery would result in 173 in-hospital deaths in the first column and 712 survivors in the second column. Considerations of this kind will be important in Chapter 3.

Structure

> Mathematical notation appears as a . . . language well adapted to its purpose, concise and precise, with rules which, unlike the rules of ordinary grammar, suffer no exception.
>
> —GEORGE POLYA[1]

A Population

The $885 = 439 + 446$ patients in the ProCESS Trial in Chapter 1 are a small, finite population, not unlike the population of people living in a small town. We could refer to the patients by name, but it is more expedient to number them: $i = 1, 2, \ldots, I = 885$. Here, i refers to an individual, and I refers to the total number of individuals, in this case $I = 885$ individuals. Referring to patients in this way has several advantages. We can refer to a particular patient, say Harry, who happens to be the seventeenth patient, $i = 17$. We can refer to a generic patient, anyone at all, as patient i with i unspecified, and then make statements that are true of each and every individual. To say that such and such is true of patient i with i unspecified is to say it is true of anyone at all: true of patient $i = 1$, true of patient $i = 2$, true of Harry, or true of $i = 17$, and so on. By replacing 885 patients by I patients, we can be faithful in describing the ProCESS Trial while also recognizing that many specifics of the ProCESS Trial are incidental—for instance, the sample size—and that what we are saying is just as true of any randomized experiment formed by flipping fair coins.

Covariates

The symbol x_i will represent the observed covariates for patient i. In Table 1.1, there are nine covariates for each patient i, and x_i records the values of these nine covariates for patient i. Several covariates are attributes that may be present or absent rather than numbers, but it is customary to record an attribute as 1 if it is present and 0 if it is absent; then, the mean of an attribute is the proportion of times the attribute is present, and 100 times this mean is the percentage, as in Table 1.1. Harry, patient $i=17$, is 52 years old, male (recorded as 1), not from a nursing home (recorded as 0), with sepsis from pneumonia (recorded as 1), without sepsis from a urinary tract infection (recorded as 0), without sepsis from an intra-abdominal infection (recorded as 0), with an Acute Physiology and Chronic Health Evaluation (APACHE) II score of 23, systolic blood pressure of 96.1, and serum lactate level of 5.3. Overall, if we compare Harry with Table 1.1, we see that Harry is little younger but also a little sicker than the population as a whole, with lower blood pressure than the average, a higher APACHE II score, and a higher level of serum lactate.

Because Harry is patient $i=17$, we have been discussing x_{17}. Here, x_{17} is a small table or array with nine numbers (Table 2.1). Each patient i of the $I=885$ patients has such a table x_i of nine numbers describing patient i. If we were so inclined, we could make a table with $I=885$ rows and nine columns, where row i describes patient i, column 1 records age, column 2 records gender with male recorded as 1, and so on. Row $i=17$ describes Harry.

The symbol u_i will represent the unobserved covariates for patient i. The structure of u_i is similar to the structure of x_i, but u_i is a covariate that we did not measure. What is in u_i? Perhaps u_i includes an indicator, 1 or 0, of a variant of a gene relevant to surviving septic shock, perhaps a gene whose

Table 2.1. The value of the nine observed covariates x_{17} for patient 17

| | Background | | | Source of sepsis | | | Physiology | | |
Patient	Age	Male	Nursing home	Pn	UTI	A	APACHE II	Systolic BP	Serum lactate
x_{17}	52	1	0	1	0	0	23	96.1	5.3

A = intra-abdominal infection; APACHE II = Acute Physiology and Chronic Health Evaluation II; BP = blood pressure; Pn = pneumonia; UTI = urinary tract infection.

importance has yet to be discovered. Perhaps u_i indicates the particular strain of bacteria responsible for the infection, including its resistance to various antibiotics. Perhaps u_i records the extent of the experience of the medical resident engaged in caring for patient i. Perhaps u_i describes the social support available to patient i. Here, u_i may record one unobserved covariate or several.[2]

Two additional comments about x_i and u_i are needed. First, covariates, whether observed like x_i or unobserved like u_i, exist in a single version, as discussed in Chapter 1. In particular, patient $i = 17$ would have the x_i given in Table 2.1 whether Harry is randomized to aggressive treatment or to less-aggressive treatment. Second, in a completely randomized trial such as the ProCESS Trial, the chance that any patient i receives aggressive treatment is the same as the chance that this patient receives less-aggressive treatment; that chance does not depend on x_i, and it does not depend on u_i. We know this because we assigned treatments by flipping a fair coin. The chance that Harry is assigned to aggressive treatment is 1/2, and it does not matter, so far as that chance goes, that Harry is 52 years old with an APACHE II score of 23. Things would be different in the absence of random assignment to treatments, but the ProCESS Trial was randomized.

Treatment Assignments

Patient i is assigned to either aggressive treatment or to less-aggressive treatment, as recorded in Z_i. If a patient is assigned to aggressive treatment, this is recorded as $Z_i = 1$. If a patient is assigned to less-aggressive treatment, this is recorded as $Z_i = 0$. Harry, patient $i = 17$, was assigned to aggressive treatment, so $Z_{17} = 1$. Later, as we move away from the ProCESS experiment, $Z_i = 1$ will signify assignment to "treatment," and $Z_i = 0$ will signify assignment to "control." Here, treatment and control are generic names for two treatment groups. It is often convenient to have a symbol for the number of people in the treated group, and the symbol m will be used. In the ProCESS Trial, $m = 439$ patients received the aggressive protocol; that is, there are $m = 439$ patients i who have $Z_i = 1$.

The ProCESS Trial assigned treatments at random by flipping a fair coin, so that Z_i is a random quantity taking the value $Z_i = 1$ with probability 1/2

and the value $Z_i = 0$ with probability 1/2. We write this as $\Pr(Z_i = 1) = 1/2 = \Pr(Z_i = 0)$. It will be convenient later on to have a symbol for the probability that patient i receives treatment, a symbol for $\Pr(Z_i = 1)$, and we will use π_i for that purpose: $\pi_i = \Pr(Z_i = 1)$. As the ProCESS Trial is a completely randomized trial, $\pi_i = 1/2$ for $i = 1, \ldots, I$ where $I = 885$. Much of the complexity of causal inference arises when π_i varies from person to person in ways we do not fully understand, but this complexity is avoided in the ProCESS Trial because treatments were assigned by flipping a fair coin, so $\pi_i = 1/2$ for $i = 1, \ldots, I$.

How should we understand these probabilities, $\pi_i = \Pr(Z_i = 1)$? In a completely randomized trial, the π_i are very simple, but they are simple in a miraculous way. The value π_{17} for Harry, patient $i = 17$, takes account of everything about Harry before his treatment assignment. It takes account of every aspect of Harry's DNA, the connections of every one of Harry's neurons, and every thought Harry ever had before his treatment assignment, including thoughts that it would take years of therapy for Harry to consciously experience. If we had wanted to use this information to predict anything else about Harry's future—not his treatment assignment but something else—then we would be in a difficult spot, because we do not have this information and we would not know what to do with it if we had it. We do know, however, that if we were somehow given all of this information about Harry's past and were asked to predict Harry's treatment assignment as accurately as possible in the best possible way, then the answer would be this: Harry is as likely to receive the aggressive protocol as he is to receive the less-aggressive protocol, $\pi_{17} = \Pr(Z_{17} = 1) = 1/2$. We know this because Harry's treatment assignment was determined by the flip of a fair coin whose heads or tails owes nothing to Harry's past.

Expressed in a different way, if we wanted to compare identical treated and control groups, then we would face an unthinkable challenge. We would have to know everything about Harry's past, and that is not possible. Even then, even if we knew everything about Harry's past, that still would not be enough. We would need to find a second person, a second Harry, with a past identical to Harry's past; we would put the first Harry in the treated group and the second Harry in the control group. However, we cannot create treated and control groups with identical pasts—it is unthinkable. It cannot be done.

We are seeing that there is an enormous asymmetry between two tasks that both refer to every aspect of Harry's past before treatment assignment. One task is to use every aspect of Harry's past to create identical treated and control groups, which cannot be done. The second task is to ensure that absolutely no aspect of Harry's past influences his treatment assignment, which is straightforward: you flip a fair coin. Happily, as seen in Chapter 3, success at the second straightforward task is all that is needed for causal inference.

Effects Caused by Treatments

Potential Outcomes under Alternative Treatments

Patient i has two potential outcomes, one if patient i is given aggressive treatment, and the other if patient i is given less-aggressive treatment. Likewise, Harry, patient $i = 17$, has two potential outcomes: one under aggressive treatment, and the other under less-aggressive treatment. Harry was assigned to aggressive treatment, $Z_{17} = 1$, so we see only his outcome under aggressive treatment. Jerzy Neyman and Donald Rubin emphasized the importance of expressing causal effects explicitly as comparisons of potential responses under competing treatments.[3] To avoid tying the notation to the details of the ProCESS Trial, the notation will refer to treatment and control, not to aggressive or less-aggressive treatment.

Here, treatment and control signify two treatments being compared in a randomized experiment. That is, treatment and control are generic names for the two treatments, with no further implications. The notation uses T for treatment and C for control. If patient i receives the aggressive protocol, the patient will exhibit response or outcome r_{Ti}, whereas if patient i receives the less-aggressive protocol, the patient will exhibit response or outcome r_{Ci}. In Table 1.2, $r_{Ti} = 1$ if patient i would die in the hospital before 60 days if the patient received the aggressive protocol, and $r_{Ti} = 0$ if the patient would be discharged from the hospital alive before 60 days or be alive in the hospital at 60 days. In parallel, $r_{Ci} = 1$ if patient i would die in the hospital before 60 days if the patient received the less-aggressive protocol, and $r_{Ci} = 0$ if the patient would be discharged from the hospital alive before 60 days or be alive in the hospital at 60 days. For Harry, patient $i = 17$, we see r_{T17} but not r_{C17} because Harry was assigned to the aggressive protocol, $Z_{17} = 1$. Poten-

tial outcomes are sometimes called counterfactuals because they describe what would have happened to Harry if, contrary to fact, Harry had received the treatment he did not receive.

To say that patient i would have died in the hospital before 60 days with less-aggressive treatment but would have survived to discharge or 60 days with aggressive treatment is to say that $r_{Ti} = 0$ but $r_{Ci} = 1$. In brief, if $(r_{Ti}, r_{Ci}) = (0,1)$, then less-aggressive treatment would cause patient i to die in the hospital.

In contrast, if $r_{Ti} = r_{Ci} = 1$, then patient i would die in the hospital before 60 days whether given aggressive or less-aggressive treatment. If $r_{Ti} = r_{Ci} = 0$, then patient i would be discharged alive before 60 days or be alive in the hospital at 60 days whether given aggressive or less-aggressive treatment. If every patient i fell into one of two categories, either $r_{Ti} = r_{Ci} = 1$ or $r_{Ti} = r_{Ci} = 0$, then it would not matter whether aggressive or less-aggressive care is given: some people live, others die, but the type of care would not change who lives and who dies.

Chapter 1 discussed Fisher's hypothesis of no difference in effects of the two treatments. Fisher's hypothesis asserts that $r_{Ti} = r_{Ci}$ for every patient i in the ProCESS Trial. We write this hypothesis as H_0: $r_{Ti} = r_{Ci}$, $i = 1, \ldots, I$, where $I = 885$ in the ProCESS Trial.

It would be reasonable to describe the pair, (r_{Ti}, r_{Ci}), as the causal effect, because it gives us everything. Sometimes, people speak of $\delta_i = r_{Ti} - r_{Ci}$ as the causal effect, although I will call this the causal effect difference. Fisher's null hypothesis says $r_{Ti} - r_{Ci} = 0$ for every i or equivalently $\delta_i = 0$ for every i. Aggressive treatment saved patient i if $\delta_i = r_{Ti} - r_{Ci} = 0 - 1 = -1$ but it caused patient i's death if $\delta_i = r_{Ti} - r_{Ci} = 1 - 0 = 1$, and so on. We see part of (r_{Ti}, r_{Ci}), either r_{Ti} or r_{Ci}, but we never see $\delta_i = r_{Ti} - r_{Ci}$.

For any one patient, we see only one of the two potential outcomes. If the patient actually receives the aggressive protocol, we see r_{Ti} but cannot see r_{Ci}. If the patient actually receives the less-aggressive protocol, we see r_{Ci} but cannot see r_{Ti}. In Table 1.2, we saw r_{Ti} for the 439 patients in the first row of Table 1.2, and we saw r_{Ci} for the 446 patients in the second row. We never see both r_{Ti} and r_{Ci} for the same patient i. As discussed in Chapter 1, this is what makes causal inference difficult: causal statements refer to r_{Ti} and r_{Ci} jointly, but we never see r_{Ti} and r_{Ci} jointly.

In a randomized experiment, we see r_{Ti} for a random half of the population, and we see r_{Ci} for the complementary random half. This will allow us

to draw inferences about causal effects on the population, $i = 1, \ldots, I,$ without being able to draw causal inferences about any one individual i.

The Observed Response

Each patient has two potential responses, r_{Ti} or r_{Ci}, and one observed response. We observe response r_{Ti} if patient i receives the aggressive protocol, that is, if $Z_i = 1$, and we observe response r_{Ci} if patient i receives the less-aggressive protocol, $Z_i = 0$. Let us write R_i for the observed response from patient i. Harry, patient $i = 17$, received aggressive treatment, $Z_{17} = 1$, so Harry's observed response is $R_{17} = r_{T17}$, and Harry's response under less-aggressive treatment, r_{C17}, is not something we can ever see. The responses in Table 1.2 are these observed responses, R_i, the responses patients exhibit under the treatments they actually received. Table 1.2 is a tabulation of two things we observe, namely, Z_i in the rows and R_i in the columns.

When we speak in English, it is natural and straightforward to say that such and such happens if such and such happens. In mathematical notation, the use of "if" is less natural. In particular, $Z_i = 1$ if patient i received aggressive treatment or $Z_i = 0$ and $1 - Z_i = 1$ if patient i received less-aggressive treatment, so we can use Z_i and $1 - Z_i$ in place of "if." This results in concise expressions that are unambiguous. In particular, another way of writing R_i is as $R_i = Z_i\, r_{Ti} + (1 - Z_i)r_{Ci}$, because this becomes $R_i = 1r_{Ti} + (1 - 1)r_{Ci} = r_{Ti}$ if $Z_i = 1$, and it becomes $R_i = 0r_{Ti} + (1 - 0)r_{Ci} = r_{Ci}$ if $Z_i = 0$. That is, the expression $R_i = Z_i\, r_{Ti} + (1 - Z_i)r_{Ci}$ equates R_i to the right potential outcome, either r_{Ti} or r_{Ci}, without lots of English saying if-this-then-do-that.

Averages in Populations and Samples

Population Size I and Sample Sizes m and $I - m$

Recall that $Z_i = 1$ if patient i is in the treated group, the aggressive protocol, and $Z_i = 0$ if patient i is in the control group, the less-aggressive protocol, and this is true for $i = 1, \ldots, I = 885$. Because of this, the number of patients in the treated group is $m = 439 = Z_1 + Z_2 + \cdots + Z_{885}$, where "..." signifies that the omitted terms Z_3 through Z_{884} are implicitly part of the sum. In

the same way that $Z_i = 1$ picks outpatient i as a member of the treated group, $1 - Z_i = 1$ picks outpatient i as a member of the control group. In parallel, there are $I - m = 885 - 439 = 446 = (1 - Z_1) + (1 - Z_2) + \cdots + (1 - Z_{885})$ patients in the control group.

The proportion of people assigned to aggressive treatment in the ProCESS Trial is $439 / 885 = 0.496$ or 49.6% or just about half the people in the two treatment groups. That is, $0.496 = 439 / 885 = m / I = (1 / I)(Z_1 + Z_2 + \cdots + Z_I)$.

Population Averages or Means $\bar{v}$

Suppose that we have some variable v_i that is defined for each person in the population, $i = 1, \ldots, I$, where $I = 885$ in the ProCESS Trial. For instance, the APACHE II score was measured for all $I = 885$ patients in Table 1.1. The mean or average of the $I = 885$ APACHE II scores is 20.7, and we often write $\bar{v}$ for a mean. The mean or average, $\bar{v}$, is the sum of the v_i divided by the number of terms in the sum, namely, I; that is, $\bar{v} = 20.7 = (1 / 885)(v_1 + v_2 + \cdots + v_{885}) = (1 / I)(v_1 + v_2 + \cdots + v_I)$.

Sample Averages or Means, $\hat{v}_T$ and $\hat{v}_C$

Table 1.1 gives the mean APACHE II score in each of the two treatment groups, the aggressive treatment group with $Z_i = 1$ and the less-aggressive treatment group with $Z_i = 0$. Each of these is a sample mean for a simple random sample from the finite population of $I = 885$ patients in the trial as a whole. We write $\hat{v}_T$ for the mean of the $m = 439$ values v_i in the treated group, the v_i for the $m = 439$ people with $Z_i = 1$. In Table 1.1, $\hat{v}_T = 20.8$ for the mean APACHE II score in the treated group. Because $Z_i v_i$ equals $1 v_i = v_i$ if i is in the treated group, and $Z_i v_i$ equals $0 v_i = 0$ if i is in the control group, it follows that we may write $\hat{v}_T$ as $\hat{v}_T = (1 / 439)(Z_1 v_1 + Z_2 v_2 + \cdots + Z_{885} v_{885}) = (1 / m)(Z_1 v_1 + Z_2 v_2 + \cdots + Z_I v_I)$. Although the sum has $I = 885$ terms, the Z_i pick out $m = 439$ individuals whose values v_i are included in the average.

In the same way, there are $I - m = 885 - 439 = 446$ patients in the control group, and they have $Z_i = 0$, so the indicator $1 - Z_i$ picks out these $I - m = 446$ patients. The mean APACHE II score in the control group is $\hat{v}_C = 20.6$. The

mean $\hat{v}_C$ of v_i in the control group is obtained as the average of v_i with $1 - Z_i = 1$, or $\hat{v}_C = (1/446) \{(1 - Z_1)v_1 + (1 - Z_2)v_2 + \cdots + (1 - Z_{885})v_{885}\} = \{1/(I - m)\} \{(1 - Z_1)v_1 + \cdots + (1 - Z_I)v_I\}$.

Expecting Covariate Balance in Randomized Experiments

So we have three means of APACHE II scores: the population mean, $\bar{v} = 20.7$; the mean in group randomly assigned to the treatment group, $\hat{v}_T = 20.8$; and the mean in the group randomly assigned to control, $\hat{v}_C = 20.6$. The three means are similar because the treated and control groups were formed at random by flipping fair coins. In Chapter 1, we interpreted this calculation as a display or a confirmation of the covariate balance that we expected random assignment to produce: we expected the two groups to be similar because they were formed at random, and we saw that they were similar. When discussing unobserved covariates in Chapter 1, we expected the two groups to be similar because they were formed at random, but we could no longer confirm our expectation by direct inspection.

The relationship among $\hat{v}_T$, $\hat{v}_C$, and $\bar{v}$ has a third use. If we could see APACHE II scores in only one of the two treatment groups, then we could use the mean in that group to estimate the population mean; that is, we could use either $\hat{v}_T = 20.8$ or $\hat{v}_C = 20.6$ to estimate $\bar{v} = 20.7$. This third use is not important for APACHE II scores because we have all of them, but it is important in causal inference.

Estimating Population Averages from the Average in a Random Sample

We know the mean of the APACHE II scores in our finite population, namely, $\bar{v} = 20.7$, because we have every one of the $I = 885$ individual APACHE II scores, so it is just a matter of arithmetic to average them. Imagine that someone had made a mistake and had remembered to record the APACHE II scores for the $m = 439$ people who received the aggressive treatment protocol, with $Z_i = 1$, but had forgotten to record the APACHE II scores for the $I - m = 446$ people who received the less-aggressive protocol, with $Z_i = 0$. Had this mistake been made, we would not be able to calculate by arithmetic the average $\bar{v}$ of all $I = 885$ APACHE II scores because we are

missing 446 of the APACHE II scores. Nonetheless, because this is a randomized experiment, the $m = 439$ people in the aggressive treatment group are a simple random sample from the $I = 885$ patients in the ProCESS Trial, so we could quite reasonably take the mean of their $m = 439$ APACHE II scores, namely, $\hat{v}_T = 20.8$, as an estimate of the population mean $\bar{v}$. We would make an error in doing this, thinking $\bar{v}$ is 20.8 when in fact $\bar{v}$ is 20.7, but it is not a big error. It is not a big error because we have a large random sample from the population. Had this same mistake been made in reverse—remembering to record the APACHE II scores for the $I - m = 446$ patients in the less-aggressive treatment group, but forgetting to record the APACHE II scores for the $m = 439$ patients in the aggressive treatment group—then we again could not calculate $\bar{v}$ by arithmetic because we are missing 439 APACHE II scores. We could, however, reasonably take the mean of the 446 scores we have, $\hat{v}_C = 20.6$, and use them to estimate $\bar{v}$. Again, we would make a small error by using this estimate, thinking $\bar{v}$ was 20.6 when in fact $\bar{v}$ is 20.7. It is not surprising that our two sample means, $\hat{v}_T = 20.8$ and $\hat{v}_C = 20.6$, each based on roughly a random half of our finite population, are quite close to the population mean, $\bar{v} = 20.7$.

In fact, we know from the theory of sampling finite populations that, under mild conditions,[4] a mean of a large random sample is a good estimate of a population mean.[5] The theory of sampling depends heavily on the use of random sampling, the use of coin flips to pick the sample: if we had $m = 439$ measurements v_i from a population of $I = 885$ measurements but did not have a random sample picked by coin flips, then we would have no reason to believe that the sample mean is close to the population mean.

As discussed in Chapter 1, this odd little story about forgotten APACHE II scores turns out to be almost identical to the way we estimate average causal effects for the $I = 885$ patients when we cannot see the causal effect for any one patient. We see survival outcomes under aggressive treatment r_{Ti} only for the $m = 439$ patients who received aggressive treatment, and we see survival outcomes under less-aggressive treatment only for the $I - m = 446$ patients who received less-aggressive treatment, but because each is a random sample from the population of $I = 885$ patients, we can use the two sample means to estimate a population mean. As with the APACHE II scores, the key element is the random assignment of treatments, which produces two random samples from the population of $I = 885$ people. Unlike the APACHE II scores, in causal inference we do not have all of the v_i's, but we do have the two sample means of the two complementary random samples.

Average Causal Effects

What Would the Mortality Rate Have Been If Everyone Received Aggressive Treatment?

Here are two interesting causal quantities that we cannot calculate from the observable data in Table 1.2. First is $\bar{r}_T = (1/885) (r_{T1} + r_{T2} + \cdots + r_{T885}) = (1/I) (r_{T1} + r_{T2} + \cdots + r_{TI})$, which is the mortality rate for all $I = 885$ patients had they all been assigned to the aggressive protocol. We cannot calculate $\bar{r}_T$ because we see r_{Ti} only for the $m = 439$ patients who received the aggressive protocol, and we are missing the r_{Ti} for the $I - m = 446$ patients who received the less-aggressive protocol. The second quantity we cannot calculate is $\bar{r}_C = (1/885) (r_{C1} + \cdots + r_{C885}) = (1/I) (r_{C1} + \cdots + r_{CI})$, which is the mortality rate for all $I = 885$ patients had they all been assigned to the less-aggressive protocol. We cannot calculate $\bar{r}_C$ because we see r_{Ci} only for the $I - m = 446$ patients who received the less-aggressive protocol, and we are missing the r_{Ci} for the $m = 439$ patients who received the aggressive protocol.

This situation is, of course, almost exactly the same as the situation with the forgotten APACHE II scores discussed previously. The mortality rates in Table 1.2 are proportions, hence averages or means of binary variables. However, the mortality rate, $\hat{r}_T = (1/m) (Z_1 r_{T1} + Z_2 r_{T2} + \cdots + Z_I r_{TI}) = 0.21$ or 21.0% in Table 1.2, for the aggressive treatment group is not calculated from all $I = 885$ patients; rather, it is calculated from just the $m = 439$ patients who were randomly assigned to aggressive treatment.[6] As with the forgotten APACHE II scores, we have reason to expect the quantity we can calculate, namely, $\hat{r}_T = 0.210$, to be a good estimate of the quantity we want but cannot calculate, namely, $\bar{r}_T$, because $\hat{r}_T$ is the mean of a random sample of $m = 439$ of the r_{Ti}'s from a population composed of all $I = 885$ r_{Ti}'s. Unlike the forgotten APACHE II scores, we cannot check our estimate, $\hat{r}_T$, against the true value $\bar{r}_T$ because we do not know the true value $\bar{r}_T$, but we have the same reasons to think $\hat{r}_T$ is a good estimate of $\bar{r}_T$ because it is based on a simple random sample.

In parallel, the mortality rate in the less-aggressive treatment group in Table 1.2 is calculated from just the $I - m = 446$ patients who were randomly assigned to the less-aggressive treatment group; specifically, the rate in Table 1.2 is $\hat{r}_C = 0.182$ or 18.2%. Again, we have reason to expect that $\hat{r}_C$,

which we can calculate, is a good estimate of $\bar{r}_C$. We cannot calculate $\bar{r}_C$ because it involves r_{Ci} that we did not observe. We expect $\hat{r}_C$ to be a good estimate of $\bar{r}_C$ because $\hat{r}_C$ is the mean of a simple random sample of $I - m = 446$ r_{Ci} from a population of $I = 885$ r_{Ci}, and $\bar{r}_C$ is the mean of these 885 r_{Ci}.

What Is the Average Difference in Mortality Caused by the Difference in Treatments?

Of course, it seems much harder, perhaps impossible, to estimate the average, $\bar{\delta}$, of the $I = 885$ causal effect differences, $\delta_i = r_{Ti} - r_{Ci}$, because we do not see a single one of these causal effect differences. If patient i is assigned to aggressive treatment, we see r_{Ti} but not r_{Ci}, so we do not see $\delta_i = r_{Ti} - r_{Ci}$, and if patient i is assigned to less-aggressive treatment we see r_{Ci} but not r_{Ti}, so we do not see $\delta_i = r_{Ti} - r_{Ci}$. How can we estimate $\bar{\delta} = (1 / I)\,(\delta_1 + \cdots + \delta_I)$ if we never see a single $\delta_i = r_{Ti} - r_{Ci}$? Actually, estimating $\bar{\delta}$ is not hard at all, even though we have not seen a single δ_i.

Let us start with the case of a population of just $I = 2$ people; this will make the arithmetic simple. Then,

$$\bar{\delta} = \frac{\delta_1 + \delta_2}{2} = \frac{(r_{T1} - r_{C1}) + (r_{T2} - r_{C2})}{2}$$

by replacing δ_i with $r_{Ti} - r_{Ci}$. So

$$\bar{\delta} = \frac{(r_{T1} + r_{T2}) - (r_{C1} + r_{C2})}{2} = \bar{r}_T - \bar{r}_C$$

by rearranging the numerator, using $\bar{r}_T = (r_{T1} + r_{T2}) / 2$ and $\bar{r}_C = (r_{C1} + r_{C2}) / 2$. In the same way, with $I = 885$ people, $\bar{\delta}$ rearranges to equal $\bar{r}_T - \bar{r}_C$. The quantity $\bar{\delta} = \bar{r}_T - \bar{r}_C$ is called the "average treatment effect."

In Table 1.2, we estimate $\bar{r}_T$ by $\hat{r}_T = 21.0\%$ and we estimate $\bar{r}_C$ by $\hat{r}_C = 18.2\%$, so we estimate $\bar{\delta} = \bar{r}_T - \bar{r}_C$ by $21.0\% - 18.2\% = 2.8\%$. The point estimate suggests aggressive treatment increased the in-hospital mortality rate by 2.8% over what it would have been with less-aggressive treatment. Of course, $\hat{r}_T$ is just an estimate of $\bar{r}_T$, and estimates are somewhat in error. Also, $\hat{r}_C$ is just an estimate of $\bar{r}_C$, so $\hat{r}_C$ too is somewhat in error. So we still have

to ask whether our *estimate* $\hat{r}_T - \hat{r}_C = 2.8\%$ of the average treatment effect $\bar{\delta}$ could really be estimating a *population effect* that equals 0. Is $\hat{r}_T - \hat{r}_C = 2.8\%$ compatible with $\bar{\delta} = 0$? Is $\hat{r}_T - \hat{r}_C = 2.8\%$ compatible with Fisher's hypothesis that $\delta_i = 0$ for $i = 1, \ldots, I$? We need to ask whether this difference, 2.8%, could be due to chance—due to the coin flips that divided the population of $I = 885$ patients into two random samples of size $m = 439$ and $I - m = 446$ patients. Could it be that $0 = \bar{\delta} = \bar{r}_T - \bar{r}_C$ but $\hat{r}_T - \hat{r}_C = 2.8\%$ because of an unlucky sequence of coin flips, Z_i, in assigning treatments? That question will be answered in Chapter 3.

If Fisher's hypothesis of no effect is true, then every causal effect difference δ_i equals zero, so the average treatment effect is zero, $\bar{\delta} = 0$. In principle, it is conceivable that $\bar{\delta} = 0$ when Fisher's hypothesis is false, because positive effects δ_i perfectly cancel negative effects δ_i so that their average is zero. Fisher's hypothesis that every δ_i equals zero is not quite the same as the hypothesis that their mean is zero, $\bar{\delta} = 0$, but the two hypotheses are closely related.[7]

Taking Stock

The Preface recalled Frederick Mosteller's remark that you can only prove causality with statistics, and Chapter 2 has begun to indicate what he meant. If patient i is assigned to the aggressive protocol, then we see whether patient i would survive with aggressive treatment, r_{Ti}, but we cannot see whether patient i would survive with the less-aggressive treatment that i did not receive because we do not see r_{Ci}. In particular, for Harry, patient $i = 17$, the coin came up heads, so Harry was assigned to aggressive treatment, $Z_{17} = 1$; fortunately Harry was discharged alive 10 days after admission, $r_{T17} = 0$. It is now and always will be a matter of speculation whether or not Harry would have died in the hospital with less-aggressive treatment—whether $r_{C17} = 1$ or $r_{C17} = 0$—so we cannot calculate $\delta_{17} = r_{T17} - r_{C17}$. It is now and always will be a matter of speculation whether aggressive treatment caused Harry's survival, whether $\delta_{17} = -1$ or $\delta_{17} = 0$. For Harry, we know that (r_{Ti}, r_{Ci}) equals either (0,1) or (0,0), but more than this we do not know. We cannot draw a causal inference about Harry.

The situation is entirely different if we randomly split $I = 885$ patients into two random samples to give aggressive treatment to one and less-aggressive treatment to the other. We still cannot see causal effects for any one of the

$I = 885$ patients. However, we have a random half of the population to estimate $\bar{r}_T$ and another random half of the population to estimate $\bar{r}_C$, so we can say quite a bit about the average effect $\bar{\delta} = \bar{r}_T - \bar{r}_C$ of the treatment on the population of $I = 885$ patients. A causal inference in a randomized experiment is a statistical inference from random sample to sampled population, just as Mosteller said.

Causal Inference in Randomized Experiments

Is No Effect Plausible?

Are Aggressive and Less-Aggressive Protocols Equally Effective?

Is it plausible, in the ProCESS Trial, that the aggressive and less-aggressive protocols were equally effective? Suppose that you were initially inclined to believe that the aggressive protocol prevented no deaths and caused no deaths among the $I = 885$ patients when contrasted with the less-aggressive protocol. That is, suppose you were inclined to believe that some patients survive septic shock and others do not, perhaps depending upon the nature of the infection and the health of the patient, but switching protocols would not change who survives and who dies. Would the trial results force you to revise this belief?

We saw in Table 1.2 that the $m = 439$ patients assigned to aggressive treatment had an in-hospital 60-day mortality rate of 21.0% and the $I - m = 446$ patients assigned to less-aggressive treatment had a mortality rate of 18.2%; however, that observation by itself does not answer the question. The two mortality rates just quoted refer to different patients, and perhaps different patients have different mortality rates simply because they are different pa-

tients, with different infections and different states of health. True, this is a randomized experiment, so the two groups are truly random samples from the population of $I = 885$ patients, and we have every reason to believe the two groups are similar as groups even though the patients within the groups are heterogeneous. Indeed, in Table 1.1, we saw the groups were similar in a few ways, albeit not identical. The question is this: Could the difference between a mortality rate of 21.0% and a mortality rate of 18.2% be due to chance, the flip of a coin that assigned one patient to the aggressive protocol and the next to the less-aggressive protocol? Or is the 2.8% difference in mortality too large to be due to chance? Is the 2.8% difference a clear sign that aggressive treatment of septic shock kills some patients? We know in Table 1.1 that the difference between 53% male in the aggressive treatment group and 56% male in the less-aggressive treatment group is due to chance, the coin flips that formed the two groups: aggressive treatment of septic shock does not change a person's gender. When an outcome (such as mortality) differs in the treated and control groups, it could be due to chance or it could be a treatment effect, but when a covariate (such as gender) differs in the treated and the control groups, then we know it is due to chance. Could the 2.8% difference in mortality be due to chance?

Expressing the same question in different words, we are asking whether mortality rates of 21.0% and 18.2% could easily arise by chance if Fisher's null hypothesis of no treatment effect were true. Here, "no effect" is shorthand for "no difference between the effects of the two treatments being compared." Fisher's hypothesis H_0 of no effect asserts that $r_{Ti} = r_{Ci}$ for all i, $i = 1, \ldots, I$, where $I = 885$ in the ProCESS Trial. This says that patient i might die or not, but the survival of patient i under aggressive treatment, r_{Ti}, is the same as survival of patient i under the less-aggressive treatment, r_{Ci}. Equivalently, Fisher's hypothesis H_0 of no effect asserts that the effect difference, $\delta_i = r_{Ti} - r_{Ci}$, is zero for every patient, $H_0 : \delta_i = 0$, $i = 1, \ldots, I$. It is perfectly conceivable that Fisher's hypothesis of no effect is true, which implies $0 = \bar{\delta} = \bar{r}_T - \bar{r}_C$, yet $\hat{r}_T - \hat{r}_C = 2.8\%$ because $\hat{r}_T$ was estimated from patients in the aggressive treatment group while $\hat{r}_C$ was estimated from different patients in the less-aggressive treatment group. If H_0 were true, could the coin flips that assigned one patient to treatment and another to control have easily produced mortality rates of 21.0% and 18.2% in the two treatment groups?

Tabular Notation

Table 3.1 repeats the data from Table 1.2 while adding the notation from Chapter 2. Table 3.1 adds one new symbol, $T = 92$, for the count in the upper left corner cell, the number of deaths in the aggressive group. Fisher used T as the test statistic, and we will follow his example.

There are two differences between the first and second row of Table 3.1. The first difference is that the first row describes one group of $m = 439$ people and the second row describes a different group of $I - m = 446$ people. These two groups of patients where formed by flipping a fair coin $I = 885$ times, so we expect the groups to be similar as groups, but they are, nonetheless, different people. The second difference between the first and second row of Table 3.1 is that people in the first row received the aggressive protocol of treatment for septic shock, while the people in the second row received the less-aggressive protocol. So there are two possible explanations for the different mortality rates in the two rows: different people, different treatments.

Stating the same issues a little more precisely, in the first row of Table 3.1, the people were assigned to aggressive treatment by a coin that came up heads, with $Z_i = 1$, and their observed responses R_i were their responses to aggressive treatment, $R_i = r_{Ti}$. For instance, for Harry, patient $i = 17$, the coin came up heads, $Z_{17} = 1$, and Harry survived, $R_{17} = 0$, so Harry is one of the 347 survivors in the first row of Table 3.1; that is, we saw $R_{17} = r_{T17}$, but we can only speculate about what would have happened, r_{C17}, to Harry had the coin come up tails and placed him in the second row. In the second row, the coin came up tails, $Z_i = 0$, and the observed response R_i is the response to less-aggressive treatment, $R_i = r_{Ci}$. In particular, the test statistic $T = 92$ is the number of deaths, $R_i = 1$, in the treated group, $Z_i = 1$, so $T = Z_1 R_1 + \cdots + Z_I R_I = Z_1 r_{T1} + \cdots + Z_I r_{TI} = m\,\hat{r}_T = 439 \times 0.210$.

Table 3.1.   In-hospital mortality in the ProCESS Trial with general notation

Treatment group	In-hospital mortality		Total	Death rate (%)
	Death $R_i = 1$	Other $R_i = 0$		
Aggressive, $Z_i = 1$	$T = 92$	347	$m = 439$	$\hat{r}_T = 21.0$
Less aggressive, $Z_i = 0$	81	365	$I - m = 446$	$\hat{r}_C = 18.2$
Total	173	712	$I = 885$	

To repeat, $\hat{r}_T = 21.0\%$ and $\hat{r}_C = 18.2\%$ can differ for two reasons: (i) they are computed from different people, and (ii) it is possible that aggressive treatment has different effects than less-aggressive treatment, $r_{Ti} \neq r_{Ci}$ for some patients i. Fisher's null hypothesis of no effect, H_0, is a denial of the second explanation: it says that $r_{Ti} = r_{Ci}$ for every patient i, so $\hat{r}_T = 21.0\%$ and $\hat{r}_C = 18.2\%$ differ solely because of how the coins picked people for the two groups. We want to ask this question: If Fisher's null hypothesis of no effect, H_0, were true, what is the chance that coin flips would, by chance alone, put $T = 92$ or more deaths in the aggressive group? That chance, that probability, is called the P-value or significance level.

The Uniformity Trial

What Is a Uniformity Trial?

In the 1920s and 1930s randomized experimentation was a new idea, and people were unsure whether it worked, how it worked, how well it worked, or how it should be implemented in practice. To find out, they ran what they called uniformity trials.[1] Randomized experimentation was initially proposed for use in agriculture, so uniformity trials took place on an experimental farm. The farm was divided into many plots, and the plots were randomized to treatment or control. In an actual experiment, not a uniformity trial, the treated plots would receive one treatment, perhaps a new fertilizer or insecticide, and the control plots would receive another treatment. In a uniformity trial, the distinction between treated and control plots was retained, but all plots were treated in the same way, with the same fertilizer and the same insecticide. Using uniformity trials, investigators learned empirically how much treated and control groups could differ when there was no treatment effect because all plots were treated in the same way.

We saw something similar in Chapter 1 when checking covariate balance in Table 1.1. We know the variables in Table 1.1 were not affected by the two protocols of care for septic shock, so we know that these differences are due to chance. The differences in Table 1.1 are not large, but we did see a 3% difference in the proportion of men, and we know this is due to chance. We are currently asking whether a 21.0% – 18.2% = 2.8% difference in mortality could be due to chance.

Table 3.1 tabulates the assigned treatment Z_i as the rows, and the observed response R_i as the columns, where in the first row R_i equals the response to treatment r_{Ti}, and in the second row R_i equals the response to control r_{Ci}. The two rows of Table 3.1 may differ for two reasons: (i) they refer to different people, and (ii) they record different things, r_{Ti} in the first row and r_{Ci} in the second. In contrast, the analogous table in a uniformity trial is simpler. Because everyone in a uniformity trial receives the same treatment, say, the control, both rows of the table analogous to Table 3.1 record the assigned treatment Z_i, and the response to control r_{Ci}. In the uniformity trial, the two rows of the table analogous to Table 3.1 differ solely because they refer to different people, because both groups of people received the same treatment.

The Uniformity Trial and Fisher's Hypothesis

The ProCESS Trial was an actual experiment, not a uniformity trial. Had it been a uniformity trial—had the groups been formed at random in the same way, but with everyone receiving the less-aggressive standard protocol of care—we would not have observed r_{Ti} for any patient i, and would instead have observed r_{Ci} in both the first and second rows of Table 3.1. In particular, the upper left corner cell of Table 3.1 would not have been $T = 92 = Z_1 r_{T1} + \cdots + Z_I r_{TI}$, rather it would have been $T^* = Z_1 r_{C1} + \cdots + Z_I r_{CI}$. In the actual ProCESS Trial, we cannot see T^* because it is the number of deaths that would have occurred among the people in the aggressive treatment group had they been treated using the less-aggressive protocol. To calculate T^*, we would have to know what would have happened, r_{C17}, to Harry, patient $i = 17$, had Harry been treated with the less-aggressive protocol, when in fact he received the aggressive protocol; so we saw r_{T17}, not r_{C17}. In a sense, T^* is precisely what we do not know: What would have happened to patients had they received the treatment they did not receive?

If Fisher's hypothesis of no effect, H_0, were true, so that $r_{Ti} = r_{Ci}$ for every patient i, $i = 1, \ldots, I = 885$, then $T = T^*$. So testing the hypothesis of no effect consists of asking whether T exhibits the sort of behavior we would expect of T^* in a uniformity trial, a trial with no treatment effect, or whether

we need to believe $r_{Ti} \neq r_{Ci}$ for some patients i if we are to make sense of the behavior of T.

The quantity $A = Z_1(r_{T1} - r_{C1}) + \cdots + Z_I(r_{TI} - r_{CI}) = Z_1\delta_1 + \cdots + Z_I\delta_I = T - T^*$ is called the attributable effect: it is the net increase in deaths in the treated group caused by receiving treatment rather than control. Again, we cannot see A because we cannot see T^*. If Fisher's hypothesis of no effect were true, $r_{Ti} = r_{Ci}$ for every patient i, $i = 1, \ldots, I = 885$, then $A = 0$.

Indeed, if Fisher's hypothesis of no effect were true, then Table 3.1 would be, for all intents and purposes, the results of a uniformity trial. Saying the same thing in different words: if $r_{Ti} = r_{Ci}$ for every patient i, $i = 1, \ldots, I = 885$, then like a uniformity trial Table 3.1 would record the treatment assignment, Z_i, in the rows and the response to control, r_{Ci}, in the columns for the simple reason r_{Ti} and r_{Ci} are the same.

At the beginning of the chapter, we asked, If Fisher's null hypothesis of no treatment effect, H_0, were true, what is the chance that coin flips would, by chance alone, put $T = 92$ or more deaths in the aggressive treatment group? We may now rephrase this question: If Table 3.1 had come from a uniformity trial, what is the chance that there would be 92 or more deaths in the aggressive treatment group?

Testing No Effect: A Small Example

Possible Treatment Assignments and Their Probabilities

Before testing the hypothesis of no effect in the ProCESS Trial with its $I = 885$ patients, it is helpful to imagine in detail a small version of the trial. To emphasize, this is an imagined trial, entirely made up to illustrate various issues in a small but extreme case. We will return to the actual ProCESS Trial after considering the imagined trial.

Imagine a randomized clinical trial with $I = 8$ patients, $i = 1, \ldots, I$, of whom $m = 4$ are picked at random for treatment, say, the aggressive protocol in the ProCESS Trial, and the remaining $I - m = 4$ patients receive the control, say, the less-aggressive protocol. Patient i has $Z_i = 1$ if assigned to the aggressive protocol and $Z_i = 0$ if assigned to the less-aggressive protocol, and $4 = m = Z_1 + \cdots + Z_8$.

Table 3.2. Abbreviated table of 70 possible treatment assignments

			Treatment indicators							
Assignment	Probability	Treated	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8
1	1/70	1, 2, 3, 4	1	1	1	1	0	0	0	0
2	1/70	1, 2, 3, 5	1	1	1	0	1	0	0	0
			⋮							
44	1/70	2, 3, 6, 8	0	1	1	0	0	1	0	1
			⋮							
70	1/70	5, 6, 7, 8	0	0	0	0	1	1	1	1

There are 70 ways to pick $m = 4$ patients for treatment from $I = 8$ patients. For instance, you could pick patients 1, 2, 3, and 4 for treatment—that is one way to pick four from eight. Or you could pick patients 1, 2, 3, and 5 for treatment—that is a second way. Continuing in that manner, one finds there are 70 ways to pick $m = 4$ patients from $I = 8$ patients.[2]

Whenever you pick four of eight patients for treatment, you set four of the Z_i equal to 1 and the remaining four to 0. If you picked patients 1, 2, 3, and 5 for treatment, you set $Z_1 = Z_2 = Z_3 = 1$, $Z_4 = 0$, $Z_5 = 1$, $Z_6 = Z_7 = Z_8 = 0$. So there are 70 ways to set four of the Z_i equal to 1 and the remaining four Z_i equal to 0.

In a randomized experiment in which $m = 4$ patients are picked at random for treatment from $I = 8$ patients, every group of four patients has the same chance of being picked for treatment—this is the essential, defining feature of a randomized experiment. As there are 70 possible ways to pick $m = 4$ patients from $I = 8$ patients, each of these 70 possible choices has probability 1/70. Generally, in a completely randomized experiment, you pick m patients at random from I patients, with every subset of m patients having the same probability. In a randomized experiment, the probabilities that govern the treatment assignments, Z_i, are known, because those probabilities were created by the experimenter in the process of randomly assigning treatments.

In highly abbreviated form, Table 3.2 summarizes the situation: the 70 possible treatment assignments, each with probability 1/70. Of course, Table 3.2 omits 66 of the 70 possible treatment assignments, as indicated by the ellipses (". . .").

The Trial Results

Using dice or random numbers from the computer, we pick one of the treatment assignments in Table 3.2, giving each possible assignment probability $1/70$. As it turns out, we picked the 44th row of Table 3.2, which put patients 2, 3, 6, and 8 in the treated group and the rest in the control group. That is, we picked $Z_1 = 0$, $Z_2 = 1$, $Z_3 = 1$, $Z_4 = 0$, $Z_5 = 0$, $Z_6 = 1$, $Z_7 = 0$, $Z_8 = 1$. There was nothing special about this treatment assignment; we picked it by luck.

Patients were then treated in the manner dictated by the random numbers, and in-hospital mortality, R_i, was recorded. For each patient i, we record $R_i = 1$ if patient i died or $R_i = 0$ if patient i survived. To make a point, the imagined trial is imagined to have extreme results for survival; specifically, $R_1 = 0$, $R_2 = 1$, $R_3 = 1$, $R_4 = 0$, $R_5 = 0$, $R_6 = 1$, $R_7 = 0$, $R_8 = 1$, yielding Table 3.3. In Table 3.3, everyone assigned to treatment died, while everyone assigned to control survived. Could Table 3.3 be due to chance? Could it be due to an unlucky treatment assignment in the absence of any difference between treatment and control? Admittedly, it is a dramatic difference, but there are only $I = 8$ patients. Could the pattern in Table 3.3 easily arise by chance alone if the treatment had no effect?

The Logic of Hypothesis Tests

To test the null hypothesis of no treatment effect, we suppose, tentatively, just for the sake of argument, that it is true. This is a part of the logic of hypothesis tests, and also sometimes a source of misunderstanding. Supposing the null hypothesis to be true in testing that hypothesis has nothing

Table 3.3. In-hospital mortality in the small hypothetical experiment

| | In-hospital mortality | | | |
| | Death | Other | | |
Treatment group	$R_i = 1$	$R_i = 0$	Total	Death rate (%)
Treated, $Z_i = 1$	$T = 4$	0	$m = 4$	$\hat{r}_T = 100$
Control, $Z_i = 0$	0	4	$I - m = 4$	$\hat{r}_C = 0$
Total	4	4	$I = 8$	

to do with believing the null hypothesis to be true—quite the contrary. Null hypotheses are not the sort of thing you believe or disbelieve, and in any case belief has no role here. In science, no one cares what you believe. Everybody has beliefs; a scientist has evidence. Rather, testing a null hypothesis asks whether the data we saw—here, Table 3.3—provide strong evidence that the null hypothesis is false, or alternatively whether the data we saw provide little guidance about whether the null hypothesis is true or false.

People often want more than this from a test of a null hypothesis. They desire to know whether the hypothesis is true or false, not whether available data constitute strong evidence against the hypothesis or else provide little guidance about its truth. This is a forlorn desire, for it is not hypothesis testing but the world itself that refuses to comply. The null hypothesis could be false in such a subtle way that no amount of investigation or data could ever reveal that it is false. Suppose that aggressive treatment were of benefit to one patient and one patient only. Suppose aggressive treatment saved Harry, patient $i = 17$, who received aggressive treatment, $Z_{17} = 1$, but aggressive treatment had no effect on anyone else, so $r_{T17} = 0$, $r_{C17} = 1$, $\delta_i = -1$, but $r_{Ti} = r_{Ci} = 0$, $\delta_i = 0$, for $i \neq 17$. Because we can never know about what Harry's fate would have been had he received less-aggressive treatment—because we saw $r_{T17} = 0$ but can never see r_{C17}—we can never know whether the treatment had no effect on anyone or whether it saved Harry but affected no one else. For physicians treating septic shock, it is most useful to know whether aggressive treatment typically benefits patients, or typically harms patients, or typically benefits patients with a particular symptom and harms patients with a different symptom; for these most useful questions, what hypothesis testing does provide can be useful if used appropriately.

The logic of hypothesis tests asks, If the null hypothesis were true, would we ever see a table like Table 3.3? Obviously, we *could* see a table like Table 3.3—it is a logical possibility—but is it wildly improbable? If the null hypothesis were true, is Table 3.3 an event like drawing a royal straight flush in five-card stud poker? The chance of royal straight flush in a fair poker game is 1.54×10^{-6} or between 1 and 2 chances in a million. If the dealer sat down, dealt the cards placing large bets, and revealed his royal straight flush, then you would have reason to doubt the premise that you are playing in a fair poker game. Analogously, if the chance of a table like Table 3.3 were one in a million when you suppose that the null hypothesis is true, then seeing Table 3.3 would give you reason to doubt that the null hypothesis is

true. Or would a table like Table 3.3 be an ordinary, everyday event were the null hypothesis is true, like getting two heads when you flip two coins? The chance of two heads when two fair coins are flipped independently is 1/4, hardly a reason to doubt the fairness of the coins.

So the question is, If the null hypothesis of no treatment effect were true in a randomized experiment with $m = 4$ people picked at random from $I = 8$ people, what is the chance that we would see the $T = 4$ or more deaths in the treated group in Table 3.3?

The Distribution of the Test Statistic T When the Null Hypothesis Is True

It is time to assemble the parts and reach a conclusion.

In our small, imagined experiment, Fisher's null hypothesis of no effect says there is no difference in the effects of the two treatments for each of the $I = 8$ patients, $r_{T1} = r_{C1}, \ldots, r_{T8} = r_{C8}$. For the purpose of testing this hypothesis, we are tentatively supposing the hypothesis is true. We randomly selected treatment assignment 44, that is, we picked $Z_1 = 0$, $Z_2 = 1$, $Z_3 = 1$, $Z_4 = 0$, $Z_5 = 0$, $Z_6 = 1$, $Z_7 = 0$, $Z_8 = 1$ and observed responses $R_1 = 0$, $R_2 = 1$, $R_3 = 1$, $R_4 = 0$, $R_5 = 0$, $R_6 = 1$, $R_7 = 0$, $R_8 = 1$, yielding Table 3.3.

Now there is a crucial step. The crucial step says, if the null hypothesis of no effect were true, then the situation would be quite simple, essentially a uniformity trial, in which we would know what would have happened in situations we did not see. Were the hypothesis true, we would know what would have happened under treatment assignments we did not see because the hypothesis stipulates that the treatment did not do anything, did not change any patient's response. Stating this more precisely, the observed response from patient i is $R_i = r_{Ti}$ if patient i received treatment, $Z_i = 1$, or is $R_i = r_{Ci}$ if patient i received control, $Z_i = 0$. However, if the null hypothesis of no effect is true, then $r_{Ti} = r_{Ci}$, so for all $I = 8$ patients, $R_i = r_{Ci}$. If the hypothesis were true, we would know from the observed data that $r_{C1} = 0$, $r_{C2} = 1$, $r_{C3} = 1$, $r_{C4} = 0$, $r_{C5} = 0$, $r_{C6} = 1$, $r_{C7} = 0$, $r_{C8} = 1$. Also, if the hypothesis were true, we would know that changing the treatments patients received would not have changed their responses. This means, if the hypothesis were true, we would know what Table 3.3 would have been under every one of the 70 treatment assignments in Table 3.2. For instance, if the null hypothesis

were true, and if we had picked the first treatment assignment in Table 3.2, with $Z_1 = 1$, $Z_2 = 1$, $Z_3 = 1$, $Z_4 = 1$, $Z_5 = 0$, $Z_6 = 0$, $Z_7 = 0$, $Z_8 = 0$, then the upper corner cell, T, in Table 3.3 would have been $r_{C1} + r_{C2} + r_{C3} + r_{C4} = 0 + 1 + 1 + 0 = 2$. In general, if the null hypothesis were true, then $T = Z_1 r_{C1} + \cdots + Z_8 r_{C8}$, so we know what T would have been in each of the 70 situations in Table 3.2. Indeed, Table 3.4 illustrates this computation, again omitting 66 of the 70 rows. In row 44 of Table 3.4, we have $T = 4$, as in Table 3.3.

In row 44 of Table 3.4, we have $T = 4$, as in Table 3.3. Either by filling in the omitted 66 rows of Table 3.4 or employing a moment's thought, one realizes that exactly one of the 70 treatment assignments yields $T = 4$, namely, assignment 44. Similarly, filling in the omitted 66 rows of Table 3.4, we find that 16 treatment assignments yield $T = 3$. Continuing to count rows, we may produce Table 3.5.

Table 3.5 is the distribution of the test statistic T when the null hypothesis of no effect is true, and it plays an important role in testing the null hypothesis. Somewhat more precisely, Table 3.5 is the distribution of T under the null hypothesis in a randomized experiment in which $m = 4$ of $I = 8$ patients were picked at random for treatment, with observed responses $R_1 = 0$, $R_2 = 1$, $R_3 = 1$, $R_4 = 0$, $R_5 = 0$, $R_6 = 1$, $R_7 = 0$, $R_8 = 1$. For instance, in this case, the largest possible value of T is $T = 4$, and only assignment 44 yields $T = 4$, and assignment 44 had probability $1/70$, so the probability that $T = 4$ is $1/70 = 0.0143$. So, under the null hypothesis, we may write $\Pr(T = 4) = 1/70$ for "the probability that T takes the value 4 is $1/70$." As $T = 4$ is the largest possible value, we also have $\Pr(T \geq 4) = 1/70$. Continuing, there are 16 rows of Table 3.4 that have $T = 3$, and each of the 16 rows has probability $1/70$, so $\Pr(T = 3) = 16/70$, and $\Pr(T \geq 3) = \Pr(T = 3) + \Pr(T = 4) = 16/70 + 1/70 = 17/70$. Continuing in this way, Table 3.5 is completed.

Table 3.5 is built from two premises and one activity. The first premise is that we assigned treatments using truly random numbers; this premise produces the probabilities, $1/70$. The second premise is that the hypothesis of no treatment effect is true, so we can deduce the value of T under all possible treatment assignments from the responses we observed. The activity is counting in a table with 70 rows, namely Table 3.4.

So what have we found? We observed $T = 4$ deaths in the treated group in Table 3.3. We found in Table 3.5 that if the treatment had no effect in this randomized experiment, then the chance that T is 4 or more is

Table 3.4. Abbreviated null distribution of T in the small hypothetical randomized experiment

			Responses under the hypothesis of no effect							
			Treatment indicators							
			$R_1=0=r_{C1}$	$R_2=1=r_{C2}$	$R_3=1=r_{C3}$	$R_4=0=r_{C4}$	$R_5=0=r_{C5}$	$R_6=1=r_{C6}$	$R_7=0=r_{C7}$	$R_8=1=r_{C8}$
Assignment	Probability	T	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8
1	1/70	2	1	1	1	1	0	0	0	0
2	1/70	2	1	1	1	0	1	0	0	0
			...							
44	1/70	4	0	1	1	0	0	1	0	1
			...							
70	1/70	2	0	0	0	0	1	1	1	1

Table 3.5.   The distribution of *T* when the null hypothesis is true

t	Possible values t of T					
	0	*1*	*2*	*3*	*4*	*Total*
Number of rows with $T = t$	1	16	36	16	1	70
$Pr(T = t)$	1/70	16/70	36/70	16/70	1/70	1
$Pr(T \geq t)$	70/70	69/70	53/70	17/70	1/70	

$Pr(T \geq 4) = 1/70 = 0.0143$, or between 1 and 2 chances in 100. So it is un-likely that the random assignment of treatments would put all of the deaths in the treated group in Table 3.3; that would happen by chance only in fewer than two experiments in 100 experiments. Concisely if imprecisely, $T \geq 4$ is too many deaths to be "due to chance," due to an unlucky choice of random numbers in assigning treatments to patients. Because we observed $T \geq 4$, and $T \geq 4$ is improbable when the null hypothesis of no effect is true, we say that Table 3.3 constitutes fairly strong evidence against the null hypothesis in our small randomized experiment.

P-Values or Significance Levels

The quantity we calculated, $Pr(T \geq 4) = 1/70$, is known as the one-sided *P*-value or one-sided significance level testing the null hypothesis of no effect. If this one-sided *P*-value is small, then a value of the test statistic *T* as large or larger than the observed value, here 4, would have been improbable if the null hypothesis were true, so a small *P*-value counts as evidence against the null hypothesis. By an arbitrary if well-entrenched convention, a *P*-value of 0.05 or less is judged small, and a *P*-value above 0.05 is judged not small; how-ever, in truth, a *P*-value of 0.049 means much the same thing as a *P*-value of 0.051, and they are both very different from 0.001 and from 0.3.

Published randomized clinical trials almost invariably report "two-sided" *P*-values, not "one-sided" *P*-values. The two-sided *P*-value here is $2 \times (1/70) = 0.02857$ and is still fairly small. The two-sided *P*-value reflects the fact that the investigators would have claimed a difference between aggressive and less-aggressive treatment had either less-aggressive treatment beaten

aggressive treatment 4 to 0, as in Table 3.3, or if less-aggressive therapy had lost to aggressive therapy 0 to 4.[3]

The two-sided *P*-value is intended to control the probability of claiming to have found something when what you found is due to chance. If either 4-to-0 or 0-to-4 results would lead you to claim to have found something, then you need to report a two-sided *P*-value. The editors of scientific journals presume that authors are eager to claim that they have found something, hence would claim either 4-to-0 or 0-to-4 as a finding, so editors typically require the reporting of two-sided *P*-values.

The coins might have fallen differently so that treatment assignment 43 rather than assignment 44 determined which patients received aggressive treatment. In row 43, not shown in Table 3.4, the value of T is 3 rather than 4. Had we seen the same patients live and die as in Table 3.4, but had we seen this with random treatment assignment 43 rather than 44, then T would have been 3 rather than 4 and the one-sided *P*-value from Table 3.5 would have been $\Pr(T \geq 3) = 17/70 = 0.243$. If Fisher's hypothesis were true in Table 3.4, then nearly 1 in 4 randomized experiments would produce $T \geq 3$. So $T \geq 4$ is somewhat strong evidence against Fisher's hypothesis of no effect, but $T \geq 3$ is a common occurrence when the hypothesis of no effect is true. To put $\Pr(T \geq 3) = 0.243$ in context, if you flipped a dime, then flipped a quarter, the probability of two heads is 0.250, and that would hardly be a reason to doubt the fairness of the coins.

Rejecting or Accepting a Null Hypothesis

It is sometimes said that a null hypothesis is "rejected" if the *P*-value is below some cutoff α, conventionally $\alpha = 0.05$, and "accepted" otherwise. If we spoke in this way, then we would say that Table 3.3 led us to reject the hypothesis of no effect at the $\alpha = 0.05$ level. This way of speaking has a useful role, but it is also the source of more than a little confusion. The useful role stems from the need to speak of making an error in hypothesis testing, the need to keep the frequency of errors under control. For instance, one error is described as "falsely rejecting a true null hypothesis." The confusion stems from the terms "accept" and "reject," which are attractively concise but unattractively inaccurate.

Scientific hypotheses are not consumer goods. At a farm stand on a country road, we might reject one apple, accept another, briefly savor the taste of the accepted apple, and soon forget both apples. Scientific hypotheses are not accepted or rejected in this sense. The published report of the ProCESS Trial is a permanent part of the scientific literature. The evidence provided by the trial about the hypotheses it entertained will be reconsidered at intervals, particularly by scientific publications that review, summarize, and attempt to reconcile many related findings in the scientific literature. These reviews may include further statistical analyses, called meta-analyses, that reevaluate data and hypotheses from the ProCESS Trial and other studies in an effort to understand consistencies and divergences among many studies.[4] The ProCESS Trial made a permanent contribution of evidence to the scientific literature, but it did not decide the final fate of the hypotheses it entertained. In this sense, the terminology of accepting or rejecting hypotheses can be misleading.

"Falsely rejecting a true null hypothesis" is an admirably concise but inaccurate way of saying "claiming to have provided relatively strong evidence against a null hypothesis when the hypothesis is in fact true." This is often called a "type-1 error" and more suggestively called a "false rejection" or sometimes a "false discovery."[5] In parallel, "falsely accepting a false null hypothesis" is a concise but quite misleading way of saying "failing to provide much evidence one way or the other about whether a null hypothesis is true." This is often called a "type-2 error" and might more suggestively be called "failing to provide much evidence." For better or worse, the universal custom is to use the concise terms and expect listeners to parse them appropriately. In this book, I will follow the custom to limited extent of speaking of a "false rejection," but I will depart from the custom in speaking of "failing to reject a null hypothesis" rather than "accepting a null hypothesis."

It is an enormous, unforgivable mistake to interpret "accepting a null hypothesis" as providing evidence that the null hypothesis is true. "Accepting a null hypothesis" means failing to provide much evidence about whether it is true or false. If the sample size is small enough, you are all but certain to "accept the null hypothesis" not because you have provided evidence in its favor, but because you have failed to provide much evidence at all. If Table 3.3 had had a sample size of $I = 2$ patients, then the test we applied to Table 3.3 could not possibly reject, at the $\alpha = 0.05$ level, the null hypothesis no matter how the results came out. Additionally, the null hypothesis may be false in

trivial ways that we can never discern—for instance, earlier in this chapter, we considered the possibility that the treatment affected Harry, patient $i = 17$, and no one else, and we saw that we could never discern this.

If you wanted to provide strong evidence that aggressive treatment and less-aggressive treatment differed negligibly in their effects, then you would not do this by "accepting the null hypothesis." You would still use the technology of hypothesis testing. However, to demonstrate that the effects are almost the same, you would need to reject every hypothesis that says they are substantially different. This activity comes in two similar but not identical wrappers. One wrapper does precisely what was requested, rejecting the hypothesis of a substantial difference in effects; this is known as equivalence testing.[6] The second, more familiar, wrapper is a confidence interval or confidence set. A 95% confidence set is the set of all hypothesis not rejected by a 0.05 level test, and if that set excludes all large differences in effects, then it provides evidence that the difference in effect is not large.[7] A later, optional section of this chapter (i.e., a section marked with an asterisk) discusses these issues in a little more detail.

If we reject a null hypothesis when an appropriate P-value is less than or equal to 0.05, then the chance that we falsely reject a true null hypothesis is at most 0.05. To see this, consider the one-sided test in Table 3.1. As seen earlier, in Table 3.1, if the null hypothesis of no effect were true, then the chance that $T \geq 4$ is $\Pr(T \geq 4) = 1/70 = 0.0143$, so the one-sided P-value is 0.0143. It follows that the chance that we falsely reject a true null hypothesis by rejecting it when $T \geq 4$ is 0.0143, which is less than or equal to 0.05.[8]

Fisher's Exact Test; the "Reasoned Basis for Inference"

The test in Table 3.5 is commonly called "Fisher's exact test" for a 2×2 table, and it was introduced in Chapter 2 of Fisher's 1935 book *Design of Experiments.* The popular name, "Fisher's exact test," does not convey the aspects of the test that interested Fisher. The word "exact" means that the distribution in Table 3.5 is exactly the distribution of the test statistic T in a completely randomized experiment when Fisher's null hypothesis is true. That is, Table 3.5 is exactly the null distribution of T, not an approximation to its null distribution. Having an exact null distribution is nice, like having a sharp crease in your trousers, but it is not one of life's larger achievements.

Approximations are widely used in statistical inference, and by and large when these approximations are well designed, they work well.

In chapter 2 of his *Design of Experiments,* Fisher spoke of randomization in experiments as the "reasoned basis for inference" in experiments. He did not mention this; he belabored it. He dealt with the exact distribution of *T* in a couple of paragraphs, then went on and on about the "reasoned basis for inference." What was he trying to convey? He was trying to convey two things, one about his null hypothesis of no effect and the other about where probability distributions come from.

Fisher's hypothesis of no treatment effect is a null hypothesis that speaks directly about the effects caused by treatments. Specifically, the hypothesis denies that there is any effect, so changing the treatment a person receives does not change the person's response. If you do not know whether or not *that* hypothesis is plausible, then you do not know much.

Commonly, statistical hypotheses refer to parameters or aspects of a convenient statistical model, and then a separate argument, not always a particularly clear or compelling argument, is invoked to connect this convenient but rather technical model to the scientific problem at hand. The arguments that connect technically convenient statistical models to important scientific questions—these connectivity arguments—are often most compelling to people who do not understand them, and least compelling to people who do. A careful, thoughtful, and instructive example—that is, a rare, good example—of such a connectivity argument is given by David Cox, contrasting Fisher's hypothesis under random assignment with various model-based hypotheses.[9]

Unlike hypotheses about parameters in statistical models, Fisher's null hypothesis is speaking directly and plainly about the most basic aspect of the scientific question that prompted the experiment: Did the treatment cause any effect? Is there any compelling evidence that the treatment is active in producing any effect at all?

Where do probability distributions come from? How do we know that the probability model used in a particular statistical analysis is the correct model, or a reasonable model, or at least not an outrageous model? These are excellent questions.[10] If these questions are little discussed, it is because most analysts do not have excellent answers. But Fisher has an excellent answer. The null distribution in Table 3.5 was derived from the coin flips that assigned patients to treatment or control—there is nothing speculative

about it. You can deny the correctness of the randomization distribution only by accusing the investigators of lying about the way they conducted their experiment and performed their analysis: if they were not lying, if they did indeed randomize, then the randomization distribution is the correct null distribution for T in Table 3.5. Where did Fisher's null distribution come from? From the coin in Fisher's hand.

Comparing Treatments in the ProCESS Trial

The ProCESS Trial Compared with the Small Example

Why did we consider a small example before considering the ProCESS Trial? In the ProCESS Trial, there were $I = 885$ patients, of whom $m = 439$ received the aggressive protocol. For the ProCESS Trial, Table 3.4 would have to be much larger. The table for the ProCESS Trial would have $I = 885$ columns, one for each patient. Also, there are 6.7×10^{264} ways to pick $m = 439$ patients from $I = 885$ patients to receive the aggressive protocol, so there are 6.7×10^{264} possible treatment assignments in the ProCESS Trial, and Table 3.4 would have 6.7×10^{264} rows, one for each possible treatment assignment. The small example with $I = 8$ patients permits inspection of details that would have been more obscure in a table with 6.7×10^{264} rows and 885 columns. Aside from its size, however, nothing new is happening in the larger table that is not already happening in Table 3.4.

Testing the Hypothesis of No Treatment Effect

In the ProCESS Trial in Table 3.1, the mortality rate in the aggressive protocol group was 21.0% and in the less-aggressive standard protocol group was 18.2%. Does this indicate the less-aggressive protocol is safer? We began by asking whether this difference in mortality rates could be due to chance, the coin flips that assigned one patient to one treatment, another patient to the other treatment. We asked whether the two treatments might have identical survival outcomes for every patient yet produce a difference in mortality rates of 2.8% simply because of the random division of one patient population

into two groups. That is, we asked whether Fisher's null hypothesis of no difference in effect between aggressive and less-aggressive protocols was plausible in light of the data in Table 3.1.

In principle, we could answer this question by going through the same steps as in the small example. The table that is analogous to Table 3.4 would now have 6.7×10^{264} rows and 885 columns. The ProCESS Investigators picked one row of this table at random, each row having probability $1 / (6.7 \times 10^{264})$, and used that row to assign treatments to the $I = 885$ patients.[11] In Table 3.1, the observed value of the test statistic is $T = 92$ deaths in the aggressive treatment group, and we want to determine the probability that $T \geq 92$ when Fisher's hypothesis of no effect is true. If the hypothesis were true, then we could append to each of the 6.7×10^{264} rows the corresponding value of T, as in Table 3.4, and with a little counting produce the relevant null distribution of T, analogous to Table 3.5. Admittedly, writing out the 6.7×10^{264} rows might require more than a few pads of paper and some careful counting, but there is no conceptual difficulty with following this approach. We would then double this one-sided P-value to obtain the two-sided P-value that is commonly reported. In practice, various technical devices yield the answer with much less computation.[12]

The one-sided P-value testing no effect turns out to be 0.1676. If, in fact, the aggressive protocol benefited no one and harmed no one when compared with the less-aggressive protocol, about 17% of the 6.7×10^{264} possible treatment assignments would have produced a difference in mortality as large or larger than the 2.8% difference seen in Table 3.1. Because the direction of the effect was not known in advance, it is customary to report the two-sided P-value as twice this one-sided P-value: $0.3352 = 2 \times 0.1676$.

The results in Table 3.1 provide no reason to abandon Fisher's hypothesis of no difference in effects of aggressive and less-aggressive protocols. It would be a commonplace event to see a difference in mortality of 2.8% or larger when there is no difference between aggressive and less-aggressive treatments: about one-third of the treatment assignments would produce such a difference by chance alone.

Hypothesis testing is asymmetric: it can provide evidence against a null hypothesis, not evidence in its favor. Hypothesis tests and P-values can reveal that there is a substantial tension or incompatibility between certain experimental results and certain hypotheses, as was true to a degree in Table 3.3. In Table 3.3, to insist that Fisher's hypothesis of no effect is true

is to insist that we were really quite unlucky, perhaps implausibly unlucky, in our random assignment of treatments. In contrast, for the actual ProCESS Trial we found that Table 3.1 would be unremarkable if Fisher's hypothesis were true. In that sense, Fisher's hypothesis is not contradicted by the results of the ProCESS Trial. Finding that Table 3.1 would be unremarkable if the hypothesis were true is finding that there is an absence of evidence against hypothesis; however, an absence of evidence against the hypothesis is not, in itself, evidence that the hypothesis is true. This is familiar from the courtroom: the absence of evidence that you are innocent is not evidence that you are guilty.

In point of fact, we can never have evidence that Fisher's hypothesis is true, but we could have strong evidence that it is false. It is easy to see why this is so. To assert that Fisher's hypothesis is true is to say that no patient benefited and no patient was harmed by the aggressive protocol, and in particular that patient $i = 17$, Harry, did not survive because he received the aggressive protocol; however, as we have seen several times, we are not, and cannot be, in a position to make such claims about individual patients.

We can, however, talk about hypotheses about all $I = 885$ patients other than Fisher's hypothesis of no effect. This is illustrated in the optional (starred) next section.

* How Large Is the Effect?

* Testing Hypotheses about Effects Other Than No Effect

So far, the focus has been on testing the null hypothesis of no effect, but there are many other hypotheses. Testing other hypotheses is a key step on the path to making inferences about the magnitude of an effect. The current section sketches a few of the ideas, leaving details to books or articles mentioned in the endnotes.[13] It is important that you know that randomization in experiments is as useful in drawing inferences about the magnitude of a treatment effect as it is in testing the hypothesis of no effect. If you are willing to accept that fact without a demonstration of it, then it would be perfectly reasonable to skip this section and continue on to Chapter 4.

The null hypothesis of no effect says $\delta_1 = 0, \ldots, \delta_I = 0$, so it specifies the value of $\delta_i = r_{Ti} - r_{Ci}$ for $I = 885$ individuals in the ProCESS Trial. In fact,

we can test any hypothesis that specifies a value δ_{0i} for δ_i for $i = 1, 2, \ldots, I$. For instance, the hypothesis would specify a value, δ_{017} for δ_{17}—the effect for Harry, patient $i = 17$. Each value δ_{0i} must be 0, 1, or −1, where $\delta_{0i} = 0$ says that aggressive treatment would not change patient i's survival, $\delta_{0i} = 1$ says aggressive would cause patient i's death, and $\delta_{0i} = -1$ says aggressive treatment would save i's life, when compared with less-aggressive treatment. Let us write H_δ for one hypothesis that specifies $I = 885$ values δ_{0i}, $i = 1, \ldots, I$, one for each patient in the ProCESS Trial. If all 885 values were $\delta_{0i} = 0$, then we would again have Fisher's hypothesis of no effect.

Because Harry received aggressive treatment, $Z_{17} = 1$, and we observed that Harry survived, $r_{T17} = 0$, we know that $\delta_{17} = r_{T17} - r_{C17}$ is either $\delta_{17} = 0 - 0 = 0$ or $\delta_{17} = 0 - 1 = -1$, depending upon Harry's unknown fate r_{C17} under less-aggressive treatment. If the hypothesis H_δ says otherwise, if it says $\delta_{017} = 1$, then H_δ is obviously false, and we may reject it with no chance of making a mistake. In parallel, H_δ may be rejected immediately if it says something obviously false about any other $I = 885$ patients. So from now on suppose H_δ does not say anything that is obviously false because we have already dealt with hypotheses that say things that are obviously false.

If H_δ were true, then we could calculate the response r_{Ci} of patient i from the hypothesized value δ_{0i}, the observed response R_i, and the treatment assignment Z_i. If H_δ were true, then $\delta_{0i} = r_{Ti} - r_{Ci}$ for each patient i. If patient i received control, $Z_i = 0$, then $R_i = r_{Ci}$, and we know r_{Ci} by direct observation. If patient i received treatment, $Z_i = 1$, then $R_i = r_{Ti}$ so that $R_i - \delta_{0i} = r_{Ti} - (r_{Ti} - r_{Ci}) = r_{Ci}$, so we can deduce r_{Ci} from the hypothesis and things we have observed. In either case, in a concise formula covering both cases, we have $r_{Ci} = R_i - Z_i \delta_{0i}$.

So if H_δ were true, we would know the response to control, r_{Ci}, for every patient. Hence, we could create a 2×2 table, recording treatment, Z_i, in the rows and response to control, r_{Ci}, in the columns. The table would resemble Table 3.1, but r_{Ci} would replace R_i in the columns. That is, if H_δ were true, we could produce the table for the uniformity trial in which everyone received less-aggressive treatment, even though that trial was never performed. In that table, the upper left corner cell would have count T^* rather than count T, and the upper right corner cell would have count $m - T^*$ rather than $m - T$ in Table 3.1. In that uniformity trial, Fisher's hypothesis of no effect is true. It follows that we may test H_δ by deducing r_{Ci} from H_δ and the observed data and testing the hypothesis of no effect in the deduced table.

If we were to count the possible hypotheses, H_δ, how many possible hypotheses would there be? There are three possible values for δ_{01} for the first patient, $i = 1$, of which one is obviously false, and there three possible values for δ_{02} for the second patient, $i = 2$, of which one is obviously false, making $4 = 2^2$ possible values for the first two patients omitting obviously false values. In the ProCESS Trial as a whole, with its $I = 885$ patients, there are $2^I = 2^{885} = 2.6 \times 10^{266}$ hypotheses H_δ that are not obviously false. Among these $2^I = 2^{885} = 2.6 \times 10^{266}$ hypotheses H_δ, there are $2^I - 1$ false hypotheses and one true hypothesis. The one true hypothesis correctly specifies the $I = 885$ values of δ_{0i}. Of course, we worry only about rejecting the one true hypothesis because that is the only false rejection of a true hypothesis; that is, we are eager to reject as many of the $2^I - 1$ false hypotheses as possible because those are correct rejections. If we tested all 2^I hypotheses H_δ, we test only one true null hypothesis, so we run the risk of falsely rejecting a true hypothesis only once.[14]

✳ Inference about Attributable Effects

We cannot see the attributable effect, $A = \sum_{i=1}^{I} Z_i \delta_i$, because we cannot see any of the causal effect differences, δ_i. However, the hypothesis H_δ specifies values δ_{0i}, so if H_δ were true, we could plug in these values and compute the attributable effect as the quantity $A_0 = \sum_{i=1}^{I} Z_i \delta_{0i}$. If H_δ were true, then A_0 would equal the attributable effect $A = T - T^* = Z_1 \delta_1 + \cdots + Z_I \delta_I$. It follows that if H_δ were true, we could compute $T^* = T - A_0$ and construct the 2×2 table from the uniformity trial. This 2×2 table would have the same second row as Table 3.1, whereas the first row would have $T^* = T - A_0$ in the upper left corner and $m - T^*$ in the upper right corner. Although there are 2^I hypotheses H_δ, all hypotheses that yield the same value of A_0 yield the same 2×2 table and the same P-value, and are accepted or rejected together.

Proceeding in this way, we find that any hypothesis H_δ that says aggressive treatment in aggregate saved 13 lives in the aggressive group, $A_0 = -13$, is rejected with a two-sided P-value ≤ 0.05.[15] In parallel, any hypothesis H_δ that says aggressive treatment in aggregate caused 35 deaths in the aggressive group, $A_0 = 35$, is rejected with a two-sided P-value ≤ 0.05.[16]

Here, 13 / 439 is about 3% of the patients who received aggressive treatment, and 35 / 439 is about 8%. It is plausible that aggressive treatment conferred no benefit and no harm on the 439 people who received it when compared to less-aggressive treatment. It is implausible that, in aggregate it saved 3% or more of the 439 patients who received it, and also implausible that it killed 8% or more because effects that large or larger have been rejected by a two-sided, 0.05 level test.[17]

Taking Stock

We began with the realization that we would never know whether patient $i = 17$, Harry, survived septic shock because he received the aggressive protocol. Harry did receive the aggressive protocol, and he did survive. We know that. We do not know, will never know, what Harry's fate would have been had he received the less-aggressive protocol. In the same way, we do not know the fate of any patient in the ProCESS Trial under the treatment that patient did not receive. As we examined the ProCESS Trial in Chapter 1, we developed an intuitive sense that we knew much more about the effects of the two protocols on population of $I = 885$ patients than we knew about any one patient in the trial. The ProCESS Trial divided a small town, a small population of $I = 885$ patients, at random into two samples. With a large random sample, we know quite a bit about the population, quite a bit about the population's survival with aggressive treatment, quite a bit about the population's survival with less-aggressive treatment, and quite a bit about the effects on the population caused by the difference between aggressive and less-aggressive treatment for septic shock.

In the current chapter, method replaced intuition. Our intuitive sense that we had learned about treatment effects in a population was replaced an objective, quantitative appraisal of the strength of the evidence and the magnitude of the effect. The test of no treatment effect depended upon our use of coin flips to assign treatments, but it entailed no speculative assumptions. Fisher described randomized treatment assignment as the "reasoned basis for inference" in experiments, and in this chapter we examined the logic of his argument.

Irrationality and Polio

Experiments on Cognition and Public Health

Chapters 1–3 considered a randomized trial in clinical medicine. Randomized experimentation is equally important in studies of behavior and cognition, in public health and public program evaluation, in the treatment of addictions, in criminology, in cognitive neuroscience that links thoughts with brain activity, and in other fields of study. In this chapter, two additional randomized experiments are considered, one concerned with irrationality and the other with public health. The chapter is optional reading as the later chapters do not refer to it.

Can a Preference Be Irrational?

A Question, Perhaps a Scientific Question

Can a preference be irrational? Is this a scientific question?

I prefer strawberry ice cream, and you prefer chocolate. Is one of us irrational? I prefer sugary soda, and you prefer imported spring water. Is rationality

an issue here? I prefer to gamble at casinos, and you prefer to save for retirement. Is my preference irrational? I prefer to have a bottle of wine as my dinner, and you prefer to have a glass of wine with dinner. Am I being irrational?

Can a preference be irrational? More than a little turns on the answer. If we take preferences as a given—preferences do not need justifications, they just are—then we are led to focus on satisfying given preferences in an efficient way. Traditionally, a course in microeconomic theory begins with a neutral stance toward given preferences, then demonstrates that markets are an efficient, and in some respects equitable, way to satisfy given preferences. To know an efficient means for satisfying given preferences is to know something useful, and microeconomic theory provides guidance. However, the task of efficiently satisfying given preferences seems less complete if preferences are often irrational. No matter how much I prefer gambling to saving for retirement, there is something to be said in favor of discouraging my preference rather than satisfying it efficiently.

How are we to decide whether a preference is irrational? If I repudiate your evaluation of my preferences, then where does that leave us? Is the existence of interminable disagreement about whether a particular preference is irrational evidence that preferences cannot be judged as rational or irrational?

I have been trying to interest you in the first question: Can a preference be irrational? But there is a second question. Is the first question a scientific question? If it is a scientific question, how can an experiment be set up to settle it?

How Could an Experiment Demonstrate That a Preference Can Be Irrational?

To convincingly demonstrate that a preference can be irrational, we would need to exhibit a situation in which a preference is (i) consequential, not trivial, (ii) widely held, (iii) yet unambiguously irrational. At first, this sounds like a tall order. At first, it sounds as if we are seeking a consequential preference that many people harbor while judging it to be unambiguously irrational. Can I recognize a preference as consequential and unambiguously

irrational, yet continue to hold that preference? Even if I could harbor a preference that I judged to be unambiguously irrational, would I not hide that fact from an investigator, and perhaps from myself, to avoid seeming foolish?

On closer examination, demonstrating (i)–(iii) does not require demonstrating that any one person holds an irrational preference. It suffices to demonstrate the existence of a population that must contain many individuals who hold an irrational preference, perhaps without being able to identify these individuals. In Chapters 1–3, we saw that we could demonstrate that a treatment had effects on a population of people without being able to identify the individuals who were affected. We did this by randomized experimentation. Perhaps we can demonstrate, again by randomized experimentation, that a form of irrationality has a hold on some members of a population of people without being able to identify the individuals who are irrational. Everyone in such an experiment might leave believing their own preferences are perfectly rational while disapproving of the manifest irrationality of other people in the experiment.

Daniel Kahneman and Amos Tversky and their colleagues conducted many such experiments.[1] Let us consider a small corner of one of these experiments.[2]

Is a Chance of Immediate Death Tolerable in Exchange for a Longer Expected Life?

A patient who is severely ill may face a choice of treatments. One treatment, perhaps a surgical treatment, places the patient at increased risk of death for a short period of time but offers much improved prospects if the patient does survive that short period of time. Another treatment, say, a form of radiation treatment, carries little immediate risk but a smaller improvement in longer-term prospects. That is a consequential choice, not a trivial one. Moreover, reasonable people might reasonably have different preferences. How can this choice be used to demonstrate that people often harbor consequential but irrational preferences?

Barbara McNeil and colleagues divided a small population of 583 individuals into two groups. The first group was given some background information and offered the following choice.

Surgery: Of 100 people having surgery, 90 live through the postoperative period, 68 are alive at the end of the first year, and 34 are alive at the end of five years.

Radiation Therapy: Of 100 people having radiation therapy, all live through the treatment, 77 are alive at the end of one year, and 22 are alive at the end of five years.[3]

In the group that was offered this choice, 18% preferred radiation therapy, and 82% preferred surgery.

The second group was offered almost exactly the same choice. The only difference was that instead of speaking of how many people would live, the description spoke of how many people would die. For instance, instead of saying that 90 / 100 people receiving surgery would survive the postoperative period, this same mortality rate was described as 10 / 100 people receiving surgery would die in the postoperative period. The mortality rates were identical all the way through the two descriptions, as was the background information. The only difference was the emphasis on living in the first description and on dying in the second description.

In the group that received the second description, 44% of people preferred radiation, and 56% of people preferred surgery. So many more people preferred radiation—a guarantee of short-term survival—when the mortality rates emphasized dying than when they emphasized living, even though the actual mortality rates are the same, and only the emphasis had changed.

In other words, we estimate a shift of 44% − 18% = 26% in population preference based on something totally inconsequential, namely, speaking of 90 / 100 surviving rather than 10 / 100 dying. This 26% shift estimates that at least 26% of the individuals in the population would change their individual preferences about a grave choice based on something inconsequential, surely a behavior that everyone in the population would regard as irrational. Yet every individual in the population can claim, without fear of contradiction, to be among the remaining 100% − 26% = 74% of the population that avoids irrationality, holding the same preference no matter how it is described.

The 583 individuals in this experiment came from three groups: patients with chronic medical problems, radiologists, and students in business school. The pattern of preference shifts was similar for all three groups.

The experiment demonstrates that a preference in a grave decision is responsive to something to which no rational preference should be responsive. If your preference in a grave matter changes in response to something you yourself regard as inconsequential, then your preference is irrational in your own eyes. Of course, you never see this about yourself, nor would anyone else. We see an irrational responsiveness to something inconsequential only for the population of 583 individuals as a whole. That is, 100% of 583 individuals are in a position to ridicule the foolish 26%. We can all agree that there are a lot of crazy people out there, but of course you and I are perfectly rational.

Let us write this argument out a little more carefully. As in Chapter 2, every individual i, $i = 1, 2, \ldots, I = 583$, has two potential preferences, r_{Ti} and r_{Ci}. Here, $r_{Ti} = 1$ if person i would prefer radiation if mortality rates were described in terms of the chance of surviving, as in the quote above, and $r_{Ti} = 0$ if person i would prefer surgery in this situation. In parallel, $r_{Ci} = 1$ if person i would prefer radiation if mortality rates were described in terms of the chance of dying, and $r_{Ci} = 0$ if person i would prefer surgery in this situation. If people were rational in their preferences about grave decisions, then their preferences would not shift based on something inconsequential. In particular, people would not change their preferences if the factual situation was unchanged but was described in a different but obviously equivalent way. You might prefer one thing, and I might prefer something else, but if we were rational, neither of us would change our preferences in response to an inconsequential change in the description of unchanged facts. That is, if the $I = 583$ people were rational, then $r_{Ti} = r_{Ci}$ for every individual i, $i = 1, 2, \ldots,$ $I = 583$. In other words, McNeil and colleagues built their experiment in such a way that rational preferences correspond with Fisher's null hypothesis of no difference in treatment effects, namely, H_0: $r_{Ti} = r_{Ci}$, $i = 1, 2, \ldots,$ $I = 583$. We saw r_{Ti} for people who heard the mortality rates described in terms of survival, the treated group with $Z_i = 1$, and we saw r_{Ci} for people who heard the mortality rates described in terms of dying, the control group with $Z_i = 0$, but we never saw both r_{Ti} and r_{Ci} for the same person i, so we never saw an individual exhibit an irrational preference, $r_{Ti} \neq r_{Ci}$.

Now the average value of r_{Ti} is $\hat{r}_T = 0.18 = 18\%$ for people who heard mortality rates described in terms of surviving, people with $Z_i = 1$. Also, the average value of r_{Ci} is $\hat{r}_C = 0.44 = 44\%$ for people who heard mortality rates described in terms of dying, people with $Z_i = 0$. So our estimate of the average treatment effect is $\hat{r}_T - \hat{r}_C = 18\% - 44\% = -26\%$. These two averages,

$\hat{r}_\mathrm{T}$ and $\hat{r}_\mathrm{C}$, describe the preferences of different people, so we need to ask whether this 26% difference could occur by chance or whether it constitutes strong evidence against Fisher's hypothesis of no effect—that is, strong evidence that some individuals i harbor irrational preferences with $r_{\mathrm{T}i} \neq r_{\mathrm{C}i}$. Reasoning as in Chapter 3, we find that the two-sided P-value testing Fisher's hypothesis is 1.4×10^{-11}, so a difference of 26% or more is very improbable with $I = 583$ people if Fisher's hypothesis were true. We conclude that Fisher's hypothesis is implausible, that many of the 583 people changed their preferences in a grave decision in response to something inconsequential.

Obviously, if you want to demonstrate that a preference can be irrational, then you cannot offer one person both choices. If you offered one person both choices, then that one person would recognize the irrationality of offering inconsistent responses to identical mortality rates described differently. However, if you offer random halves of a single population one description or the other, then you can demonstrate that the population must contain individuals who would change their preferences in response to something inconsequential. This is true even though neither the investigator nor the 583 experimental subjects knows whose preferences are consistent and whose are irrational.

This experiment and many similar experiments demonstrate that people often arrive at consequential preferences in an irrational way. However, these experiments also exhibit forms of irrationality that are fragile. Presumably, few people, perhaps no one, would have $r_{\mathrm{T}i} \neq r_{\mathrm{C}i}$ if people were forced to rewrite the description they were given in its equivalent but alternative form. The irrationality exhibited in this experiment might vanish if people were asked to consider the matter more carefully, more thoughtfully, from more than one perspective. The exhibited irrationality might vanish if people were encouraged and guided to think rationally about their preferences.

Randomized Evaluation of the Salk Poliomyelitis Vaccine

The Randomized Trial

In 1954, more than 400,000 U.S. children participated in a randomized trial to evaluate the safety and effectiveness of Jonas Salk's vaccine to prevent

polio. Then, as now, people hoped that vaccines would eradicate major viral diseases, but they also worried about the safety of vaccines. For practical purposes, the Salk vaccine trial settled the question of safety and effectiveness: the effectiveness of the vaccine far outweighed reasonable concerns about its safety.[4] Lincoln Moses remarked about the speed with which the randomized trial decided a major issue of public policy: in contrast with most questions in public health, a trial run in 1954 "settled the question once and for all in . . . 1954."[5]

In states that participated in the randomized trial, no one was forced to participate. In total, 338,778 eligible children did not participate. Among those who participated, 200,745 were randomly assigned to the vaccine, and 201,229 were randomly assigned to a placebo consisting of a saltwater solution. Looking back, using the knowledge gleaned from the trial and subsequent developments, we now know that children receiving placebo received the inferior treatment. At the time the trial was conducted, no one knew whether the vaccine actually worked, whether it had harmful side effects. No one knew whether the vaccine would do more good than harm. The opportunity to receive the vaccine was often turned down, a sign that many parents worried that the vaccine posed risks to their children.[6]

Although the randomized trial had more than 400,000 children, there were only 148 cases of paralytic polio among these 400,000 children, most cases occurring in the unvaccinated group. In the randomized trial, vaccinated children had paralytic polio at a rate of 16 per 100,000 children, while unvaccinated children had paralytic polio at a rate of 57 per 100,000. Judged using the method in Chapter 3, this difference could not reasonably be attributed to chance, to the coin flips that assigned one child to vaccine, another to placebo. Use of the vaccine was estimated to prevent more that 70% of cases of paralytic polio ($16/57 = 0.28$). The randomized trial was one important step in the near-eradication of polio worldwide.

A Parallel Observational Study

The Salk vaccine was a triumph of public health, but it was also Plan B. Originally, the evaluation of the vaccine was to have been a nonrandomized or observational study. Plan A had been to give the vaccine to children in the

second grade and to use the first and third grades as unvaccinated controls. Do you see any problems with using the first and third grade as controls?

An initial thought is that the first and third grade sound like reasonable controls. Most children in the first grade will be in second grade next year, and most children in the third grade were in the second grade last year. As controls, how bad could that be?

In fact, many states evaluated the Salk vaccine using Plan A, as an observational study of different grades, not as a randomized experiment. So we will get to see how the results compare under Plan A and Plan B. However, the health departments in many states refused to participate in Plan A, claiming that the results of Plan A would be less convincing, perhaps unconvincing, and that plan A offered no compensating advantages. In particular, these states said, If it was unethical to give the vaccine to some children and deny it to others, then it did not matter whether this division of children was based on grade in school or a lottery created by randomization.

I asked you whether you saw any problems with using the first and third grade as controls. The states that rejected Plan A emphasized two problems. Let us consider the obvious concern first. Plan B, the randomized trial, used a placebo, so when polio was diagnosed, a physician did not know whether the child had been vaccinated. Under Plan B, there was no way that the opinions of physicians about vaccination could tilt the results one way or the other.[7] You cannot hide the identity of the treatment if the control group consists of the first and third grades.

The second problem is less obvious, but perhaps more serious in this context. Under both Plan A and Plan B, no one could be forced to take the vaccine, and many parents declined it for their children. In the randomized trial, Plan B, parents who declined to participate did so before randomization, before the treated and control groups were formed. This made the trial somewhat unrepresentative of the U.S. states that participated in the trial. Parents who volunteered to participate had higher incomes and more education than those who declined to participate.[8] Mothers with low socioeconomic status were much more likely than other mothers to believe that "shots were unsafe," and were less likely to have initially learned about the trial from newspapers.[9] However, among the volunteers, the trial was truly randomized, and the treated and control groups were comparable before treatment, in particular comparable in terms of income and education and other covariates that were not measured. So volunteering and self-selection in

the randomized trial did nothing to create bias in the comparison of treated and control groups.

The situation is different under Plan A, the observational study. Some parents in the second grade refused to participate. Under Plan A, 221,998 children in states following Plan A were vaccinated, but 123,605 children in the second grade were not vaccinated. In contrast, the entire first and third grades became unvaccinated controls.[10] As a result, volunteering and self-selection altered or shaped the vaccinated group, but did not alter the control group. Even if the first and third grades were similar to the second grade, the volunteers in the second grade might differ from the entire first and third grades. To emphasize, volunteers in the second grade were unrepresentative of the U.S. states that followed Plan A and also not comparable to the controls in those states.

The vaccine appeared to be effective in the observational study, but perhaps slightly less effective than in the experiment, with 17 cases of paralytic polio per 100,000 among vaccinated second graders and 46 cases per 100,000 in the first and third grades. Here, $17/46$ is 0.37 for the observational study, in comparison with $16/57 = 0.28$ in the randomized trial.[11]

More worrisome is the finding, in both the trial and the observational study, that the rate of paralytic polio for people who refused to participate was about 35 cases per 100,000, higher than the vaccinated groups but lower than the control groups. Under Plan A, this is the rate for second graders who were not inoculated. Under Plan B, it is the rate for children who did not participate in the randomized trial.[12] In the randomized trial, it is clear that the vaccine caused a reduction in polio for children who participated in the trial, those children being unrepresentative of children in their states. That is, in the randomized trial, the vaccine clearly prevented some cases of polio among participants in the trial. The results for Plan A, the observational study, lack a similarly clear interpretation. In the observational study, the process of volunteering or refusing removed some individuals from the treated group, but did not remove anyone from the control groups. A direct comparison of volunteers from the second grade with everyone from the first and third grade may not estimate the effect of the vaccine, because those who refused the vaccine had different rates of polio from volunteers. Meier writes, "The volunteer effect . . . is clearly evident . . . Were the [Plan A] information alone available, considerable doubt would have remained about the proper interpretation of the results."[13]

The randomized trial firmly established the efficacy of the Salk vaccine in preventing polio, and it was an important step in the near-eradication of the disease. The parallel observational study was less compelling: the estimated effect appeared smaller, and there was ambiguous evidence of bias from self-selection.

What can be learned from observational studies about the effects caused by treatments? To this topic, we now turn.

Part II

OBSERVATIONAL STUDIES

Between Observational Studies and Experiments

The experimental habit of mind.
—JOHN DEWEY[1]

The Distinction between Experiments and Observational Studies

In 1965, shortly after serving as a member of the panel that wrote the 1964 U.S. Surgeon General's Report on *Smoking and Health,* William G. Cochran defined an observational study as an empiric investigation in which "the objective is to elucidate cause-and-effect relationships . . . [in which] it is not feasible to use controlled experimentation, in the sense of being able to impose the procedures or treatments whose effects it is desired to discover, or to assign subjects at random to different procedures."[2]

In a completely randomized experiment like the ProCESS Trial in Chapter 1, a coin is flipped repeatedly, assigning one person to treatment, another to control. As in a fair lottery, everyone has the same chance: the coin flips favor no one. Rich or poor, wise or foolish, everyone suffers a fate determined by luck. The central problem in an observational study—the problem that defines the distinction between a randomized experiment and an observational study—is that treatments are not assigned to subjects at random. In studies of the effects caused by smoking cigarettes, treatments—smoking or

not—are not assigned at random. In the United States in 2016, the poor are far more likely than the rich to smoke cigarettes, as the foolish are more likely to smoke than the wise. If poverty and foolish behavior have consequences for health besides increased smoking, an investigator will need to take care and exert effort to isolate the effects actually caused by smoking.

A Step beyond the Completely Randomized Experiment

An Experiment with One of the Problems of an Observational Study

There is a peculiar type of randomized experiment, rarely performed, that aids in understanding observational studies. This experiment is not a faithful description of an observational study but rather describes a much simpler and more manageable situation. We understand observational studies by understanding how they differ from this much simpler and more manageable situation.

Table 5.1 is a simulated example. In the table there are four strata defined by measured covariates, x_i, specifically age and gender. Strata are simply groups of people who are similar in terms of some measured covariates. Where Table 1.2 is called a 2×2 table because it has two row categories and two column categories, Table 5.1 is called a $2 \times 2 \times 4$ table for its two row categories, two column categories and four strata. In the first stratum, there are $100,000 = 79,828 + 20,172$ older men, whereas in the second stratum there are $100,000 = 79,779 + 20,221$ older women, and so on. For older men and for older women, the chance of receiving treatment is 0.8; that is, on average, 80% receive treatment, and 20% receive control. You could assign treatments at random in this way by rolling a die, assigning a person to treatment if a 1, 2, 3, or 4 occurred, assigning a person to control if a 5 occurred, and rolling the die again whenever a 6 occurred. For younger men and younger women, the chance of receiving treatment is 0.2; that is, on average, 20% receive treatment, and 80% receive control.

Expressed in the notation from Chapter 2, the $I = 400,000$ people in Table 5.1 differ in their probabilities of treatment, $\pi_i = \Pr(Z_i = 1)$, but the differences have a very simple, very controlled form. There are four possible

Table 5.1. A small simulated example, with randomized treatment assignment inside each of four strata, and with no treatment effect

		Stratum 1: Older men		
Group	*Dead*	*Alive*	*Total*	*Morality rate (%)*
Treated	31,868	47,960	79,828	39.9
Control	8,132	12,040	20,172	40.3
		Stratum 2: Older women		
Group	*Dead*	*Alive*	*Total*	*Morality rate (%)*
Treated	23,983	55,796	79,779	30.1
Control	6,017	14,204	20,221	29.8
		Stratum 3: Younger men		
Group	*Dead*	*Alive*	*Total*	*Morality rate (%)*
Treated	3,993	16,028	20,021	19.9
Control	16,007	63,972	79,979	20.0
		Stratum 3: Younger women		
Group	*Dead*	*Alive*	*Total*	*Morality rate (%)*
Treated	2,021	17,777	19,798	10.2
Control	7,979	72,223	80,202	9.9

values of the observed covariate, x_i: one for an older man, another for an older woman, a third for a younger man, a fourth for a younger woman. In Table 5.1, we have $\pi_i = 0.8$ for every person i whose x_i indicates an older man or woman, and $\pi_i = 0.2$ for every person i whose x_i indicates a younger man or woman. The key element is that whenever two people, say, person i and person j, look the same in terms of their observed covariate—whenever $x_i = x_j$—then they have the same probability of treatment, $\pi_i = \pi_j$. Person i and person j may differ before treatment in all sorts of important ways—they may differ in terms of unmeasured covariates, $u_i \neq u_j$—but in Table 5.1 these unobserved differences do not affect their probabilities of treatment. In Table 5.1, people who look the same in terms of measured covariates x_i are the same in one key respect: they have the same probabilities of treatment π_i.

In Table 5.1, there is no treatment effect. Some people survive, others die, but receiving treatment rather than control does not change who survives or who dies. Fisher's hypothesis of no treatment effect is true in Table 5.1;

we know this because Table 5.1 is produced by a simulation. For older men, 40% die; for older women, 30% die; for younger men, 20% die; for younger women 10% die. We know all this because the example was simulated, generated on the computer. We know that the difference between 39.9% mortality for older treated men and 40.3% mortality for older control men is just a consequence of the coin flips that assigned some older men to treatment and other older men to control. The same is true of the difference between 30.1% mortality for older treated women and 29.8% mortality for older control women. A simulated example is helpful because we know exactly what is going on—we created what is going on—and we can then compare what various analyses tell us, which analyses are reliable, which are not. Much of the mathematical theory of statistics proves theorems that tell us how simulated examples would work out; that is, the theorems describe what specific statistical methods would do if applied to data generated by specific probability models. The theorems are better than simulations because they tell us about all cases, not very particular cases like Table 5.1, and they do not depend on luck.

Table 5.1 exhibits a highly controlled departure from a completely randomized experiment. In each stratum, there is a randomized experiment: there is a perfect lottery in each stratum. The chance of a winning ticket—assignment to treatment—in stratum 1, older men, is 0.8, as in stratum 2, older women, but it is a perfectly equitable lottery for older men and a perfectly equitable lottery for older women. The chance of a winning ticket is 0.2 for younger men in stratum 3, and is also 0.2 for younger women in stratum 4. Inside stratum 3, the lottery is perfectly equitable, as it is inside stratum 4. In aggregate, however, the lottery in Table 5.1 is highly inequitable: older folks are much more likely to receive treatment than younger folks, regardless of gender.

Because each stratum on its own is a randomized experiment, we could analyze four separate randomized experiments, just as we analyzed the Pro-CESS Trial in Chapter 3. If we did this, we would find no indication of a treatment effect in stratum 1, no indication in stratum 2, and so on. The two-sided P-value testing no treatment effect is 0.31 in stratum 1 for older men, so a difference in mortality rates as large or larger than the observed 39.9% and 40.3% would happen in almost 1/3 of experiments simply because of how the coin flips turn out. In parallel, the P-value for older women is 0.40, for younger men is 0.83, and for younger women is 0.28. Viewed as four separate randomized experiments, there is no sign of a treatment effect,

Table 5.2. The four strata from Table 5.1 collapsed, leading to the false impression of a treatment effect

		Merged table		
Group	*Dead*	*Alive*	*Total*	*Morality rate (%)*
Treated	61,865	137,561	199,426	31.0
Control	38,135	162,439	200,574	19.0

which is consistent with the fact that Table 5.1 is a simulated study with no actual effect.

What would happen if we had analyzed Table 5.1 as a completely randomized experiment? In Chapter 3, in the analysis of data from the Pro-CESS Trial we might have separated people by age and gender, but instead we analyzed everyone in a single table. What would happen if we analyzed Table 5.1 in the same way?

Table 5.2 is obtained from Table 5.1 simply by ignoring age and gender. For example, the 61,865 people classified as dead after treatment in Table 5.2 result from summing the corresponding four counts in Table 5.1: 61,865 = 31,868 + 23,983 + 3,993 + 2,021. The other counts in Table 5.2 are obtained in the same way. Table 5.2 is often said to be a marginal table of Table 5.1 or a collapsed table, but really we are counting the same people, ignoring the distinctions of age and gender.

Table 5.2 yields a strikingly different impression. In this table the treated group has a mortality rate of 31.0% while the control group has a mortality rate of 19.0%. The impression is that the treatment killed 12% = 31% – 19.0% of the treated group. The *P*-value for Table 5.2 testing no effect, computed as in Chapter 3, is 2.2×10^{-16}. It is virtually inconceivable that Table 5.2 would arise by chance in a completely randomized experiment with no treatment effect. The impression from this table is entirely mistaken. We analyzed Table 5.2 as if it were from a completely randomized experiment, but it is not: it is from a peculiar randomized experiment, in which the old are likely to receive treatment and the young are likely to receive the control.

Why did this happen? Why do Tables 5.1 and 5.2 give such different impressions based on the same data? As we saw in Table 5.1, the old were more likely to die, and they were more likely to receive treatment. In Table 5.2 the treated group consists mostly of older folks who were more likely to die, and the control group consists mostly of younger folks who were less likely

to die. No surprise then: the mortality rate is higher in the treated group than in the control group, even though the treatment had no effect.

Simpson's "Paradox"

Is the relationship between Tables 5.1 and 5.2 something rare and unusual? Or is it something commonplace? It is commonplace. If you looked at a 2 × 2 table recording mortality in the United States over one year and receipt of Social Security benefits during that year, you would find a strong relationship: people who receive benefits are much more likely to die. Does this mean that the Social Security program is a menace to public health? No, it means that most people receive Social Security benefits at older ages, and most people die at older ages, so Social Security recipients die at higher rates. Many genuine relationships are not causal relationships.

Would the relationship between receiving Social Security benefits and dying vanish if we compared people of the same age—that is, if we disaggregated the 2 × 2 table to produce a table similar to Table 5.1 with separate 2 × 2 tables for each age? Perhaps, or perhaps not. Some people take reduced Social Security benefits at the earliest possible age, 62, whereas others defer the start of benefits to age 70 to receive larger benefit amounts. Suppose that you compared two people, both aged 65, one receiving Social Security benefits and the other deferring the start of benefits to a later age. It would not be surprising if the person who deferred benefits is still working and in better health, while the person who is receiving benefits has retired and perhaps is in worse health. For this reason or some other reason receiving Social Security benefits at age 65 might be related to the risk of death at age 65, so the disaggregated table might continue to indicate a relationship. Again, many genuine relationships are not causal relationships.

The pattern in Tables 5.1 and 5.2 is sometimes called Simpson's paradox, but there is nothing paradoxical about these two tables, and Simpson did not use the term "paradox."[3] Rather than "Simpson's paradox," the pattern in Tables 5.1 and 5.2 might more accurately be called "Simpson's irritating reminder" because the relationship between these tables is commonplace and perfectly intelligible, just disappointing. Simpson's irritating reminder is disappointing in the same way that the reminder "Everyone is mortal" is disappointing: no one doubts its truth or finds it puzzling, but we keep trying

to put the matter out of mind. The disappointment is that a table like Table 5.2 does not have an obvious interpretation except in a completely randomized trial: even the limited, tightly controlled, departure from complete randomization in Table 5.1 suffices to render Table 5.2 opaque and obscure.

Direct Adjustment: A Method for Estimating Average Treatment Effects

In Table 5.1, direct adjustment means estimating the average treatment effect four times, once for each stratum, then combining the four estimates with suitable weights.[4] This is an attractive procedure, but before considering why it is attractive, let us be clear about what the procedure does. Because each stratum in Table 5.1 has 100,000 people, and the entire table has 400,000 people, it is natural to give equal weights to the four strata, that is, weight $1/4 = 100,000/400,000$. Using these weights, direct adjustment yields the estimate

$$0.025 = (39.9 - 40.3)/4 + (30.1 - 29.8)/4 + (19.9 - 20.0)/4$$
$$+ (10.2 - 9.9)/4.$$

Remember that these are percentages, so 0.025 is very small, much less than 1%. Because the estimated effect of treatment is small within each of the four strata in Table 5.1, their weighted combination is also small. Where Table 5.2 incorrectly suggested an increase in mortality of 12.0% caused by treatment, direct adjustment gave an estimated effect that is 0.025 or less than three one-hundredths of 1%—essentially no effect at all. Table 5.1 is a simulated experiment in which there is truly no effect, so it is satisfying to have an estimate that reports a negligible effect, like 0.025%, and unsatisfying to have an estimate of 12% based on Table 5.2.

In the calculation just given, the four stratum specific effects received weights proportional to the total number of people in the stratum. In Table 5.1, every stratum is the same size, so these weights are equal. Unequal weights that are proportional to the total size of the stratum are natural in many if not most cases. This means that a stratum that represents a small part of the population gets a small weight and a stratum that represents a large part of the population gets a large weight. If the strata were U.S. states,

California with its enormous population would receive a much larger weight than Wyoming with its much smaller population. These weights attempt to describe the population as a whole.

Direct adjustment with stratum total weights is an attractive procedure in the following specific sense. Had we performed the peculiar randomized experiment in Table 5.1 with 0.8 probabilities of treatment for older folks and 0.2 probabilities of treatment for younger folks, then direct adjustment with stratum total weights would yield a good estimate of the average treatment effect, $\bar{\delta} = \bar{r}_T - \bar{r}_C$, defined in Chapter 2. Recall from Chapter 2 that the difference between the mean response in the treated group and the mean response in the control group yielded a good estimate of the average treatment effect in a completely randomized trial such as the ProCESS Trial. We saw that this same approach applied to Table 5.2 yielded a seriously misleading estimate in the experiment in Table 5.1 because it ignored the fact that older folks were more likely than younger folks to receive treatment. Direct adjustment fixes the problem: the estimates of the average effect within each of the four strata are individually good estimates for the individual strata because each of the four strata is a completely randomized experiment on its own, so the reasoning in Chapter 2 therefore applies. By weighting stratum-specific estimates proportionally to the stratum sizes, direct adjustment produces a good estimate of the average effect of the treatment on all the people in the finite population in Table 5.1.[5]

In brief, we can estimate the average treatment effect in the peculiar randomized experiment with 400,000 people in Table 5.1, but we cannot do this using the difference in rates in Table 5.2. To estimate the average treatment effect in Table 5.1, we need to combine the stratum specific estimates of average effect, and the combination needs to give weight to a stratum in proportion to its size.[6]

A Single Hypothesis Test Combining Stratum-Specific Results

Direct adjustment in Table 5.1 produced a single estimate of the average treatment effect by combining four stratum-specific estimates. Something similar can be done in testing the null hypothesis of no effect. For all of Table 5.1, we can have one test of the hypothesis of no effect.

Table 5.1 has 400,000 people in just four strata. Previously, we tested the null hypothesis of no effect four times, once in each of the four strata. That is not a good strategy in general, particularly in smaller studies with many more strata. With fewer people and more strata, that strategy would do many tests on tiny slivers of data. A better, commonly employed strategy uses the fact that each of the four strata in Table 5.1 is a randomized experiment on its own.

The standard test of the null hypothesis of no effect in Table 5.1 uses the total number of deaths in the treated group, $61,865 = 31,868 + 23,983 + 3,993 + 2,021$, a number that also appears in Table 5.2. However, 61,865 is not compared to the randomization distribution for Table 5.2 from Chapter 3 because that distribution is not applicable to Table 5.2. The test in Chapter 3 used the one 2×2 summary table from a completely randomized experiment in which everyone had the same chance of receiving treatment, but in Tables 5.1 and 5.2 the older folks were much more likely to receive treatment than the younger folks.

We would like to compare the total number of deaths in the treated group, 61,865, to its distribution for the experiment we actually performed in Table 5.1. The experiment we actually performed was composed of four separate, unrelated randomized experiments, stacked one on top of another, with different treatment assignment probabilities in different experiments. We need to take all of this into account. If the treatment had no effect in Table 5.1, we still expect to see more deaths in the treated group because deaths are common in strata 1 and 2, and most people receive treatment in strata 1 and 2. Given the way the study was done, even if the treatment had no effect it would not be surprising if most deaths occurred in the treated group.

In stratum 1 of Table 5.1, for older men the overall death rate is $40\% = 0.4 = (31,868 + 8,132) / 100,000$. Given this, if the treatment had no effect and 79,828 men are picked at random to receive treatment, we expect 40% of them to die, so we expect stratum 1 to contribute $0.4 \times 79,828 = 31,931.2$ treated deaths, close to the observed value of 31,868. In stratum 2 of Table 5.1, for older women the overall death rate is $30\% = 0.3 = (23,983 + 6,017) / 100,000$. Given this, if the treatment had no effect and 79,779 women are picked at random to receive treatment, we expect 30% of them to die, so we expect stratum 2 to contribute $0.3 \times 79,779 = 23,933.7$ treated deaths, close to the observed value of 23,983. Continuing in this way, if the treatment had no

effect, we expect $0.2 \times 20{,}021 = 4{,}004.2$ treated deaths among the younger men in stratum 3, close to the observed value of 3,993; and we expect $0.1 \times 19{,}798 = 1{,}979.8$ treated deaths among younger women in stratum 4, close to the observed value of 2,021. So, in aggregate, we expect $31{,}931.2 + 23{,}933.7 + 4{,}004.2 + 1{,}979.8 = 61{,}848.9$ treated deaths, close to the observed value of 61,865 in Table 5.2.

If the treatment had no effect, would the difference between the treated deaths we observed, 61,865, and the treated deaths we expected, 61,848.9, be a difference so large as to be judged improbable? We observed $61{,}865 - 61{,}848.9 = 16.1$ more deaths in the treated group than we expected if the treatment had no effect. Should we take 16.1 more deaths than expected as evidence that the treatment caused some deaths? Reasoning much as we did in Chapter 3, listing all the possible treatment assignments in Table 5, of which there are quite a few, we could determine the probability that the absolute difference between observed and expected deaths in the treated group would exceed the value we saw, namely, $16.1 = |61{,}865 - 61{,}848.9|$ deaths. That probability, the two-sided *P*-value, turns out to be 0.88. If there had been no treatment effect in the experiment in Table 5.1, then with probability 0.88 the coin flips would produce a difference as large or larger than the difference we observed, so Table 5.1 provides no indication of a treatment effect. The standard test we have been discussing is most commonly called the Mantel-Haenszel test.[7]

In brief, we can test Fisher's null hypothesis of no treatment effect in the peculiar randomized experiment in Table 5.1, but we need to do the test using the randomization that actually created Table 5.1.

Is It Important to Know the Treatment Assignment Probabilities?

As noted previously, within each stratum in Table 5.1 everyone had the same chance of receiving treatment, but that chance varied from one stratum to another. The older men and older women received treatment based on independent coin flips that came up heads with probability 0.8, while the younger men and younger women received treatment based on independent coin flips that came up heads with probability 0.2. Is it important to know the two treatment assignment probabilities, 0.8 and 0.2? Or is it enough

to know that Table 5.1 consists of four separate, completely randomized experiments?

Review the calculations done so far in Chapter 5 and you will discover that no calculation made use of the two probabilities, 0.8 and 0.2. Had someone performed the experiment in Table 5.1 just as it was performed but neglected to tell us the two probabilities, 0.8 and 0.2, we would have done exactly the same analyses, with exactly the same results.

To say this a little more precisely, in the first stratum of Table 5.1 we made extensive use of the fact that the 79,828 older men who received treatment were picked at random from the 100,000 older men in this stratum. That is, we used the fact that every sample of 79,828 older men had the same chance of treatment as every other sample of 79,828 older men, but we never used the coin's actual probability of 0.8. This was true of the four stratum-specific tests of no effect, following the pattern in Chapter 3 four times, but it was also true of the Mantel-Haenszel test that combined the four stratum-specific comparisons, and was true yet again of direct adjustment.

The peculiar randomized experiment in Table 5.1 had two attributes, one of which turns out to be critically important and the other entirely unimportant. It was important to know that each stratum was a completely randomized experiment, and that everyone in the same stratum had the same probability of treatment. It was unimportant to know what those probabilities were.[8]

What Do We Learn from This Peculiar Randomized Experiment?

Donald Rubin used the phrase "randomization on the basis of a covariate" to describe an experiment like the one in Table 5.1.[9] In such a randomized experiment, there is a known, observed covariate x_i such that the coin flips that assign individuals to treatment or control have probabilities that change with x_i. In Table 5.1, the older folks had a probability of 0.8 of treatment, while the younger folks had a probability of 0.2 of treatment. For the older folks there was an equitable randomized experiment, and for the younger folks there was an equitable randomized experiment. The inequities, the problems, occurred when the experiment for older folks was merged with the experiment for younger folks.

Randomization on the basis of a covariate is rarely performed; rather, it is an aid to understanding observational studies. With randomization on the basis of a covariate, methods for a completely randomized experiment break down; however, with various repairs and adjustments, inference about effects caused by treatments remains possible and almost straightforward. Direct adjustment yields a good estimate of the average treatment effect. The Mantel-Haenszel test accounts for treatment assignment probabilities that change from one stratum to another. Situations similar to Table 5.1 but with minor added complexities can typically be handled by methods that are, conceptually, little more than minor adjustments to these methods.[10]

In observational studies, investigators often use methods such as direct adjustment or the Mantel-Haenszel test. These methods would work if treatment assignments had been randomized on the basis of a covariate; otherwise, they might easily give the wrong inference about causal effects. Implicitly, in an observational study an investigator who offers causal conclusions based on such adjustments is suggesting that the observational study is analogous to an experiment in which treatment assignment probabilities vary from person to person only to the extent that observed covariates x_i vary from person to person. Implicitly, the critic of an observational study is suggesting the study is not, after all, analogous to a study that has been randomized on the basis of a covariate. Who is correct? Later chapters will discuss tools to aid in answering this question.

Had Table 5.1 arisen in an observational study, it would be immediately obvious that older folks are more likely to receive treatment—that is visible in Table 5.1—and immediately obvious that Table 5.2 is seriously misleading. It would be immediately obvious that you cannot apply to Table 5.2 the methods from Chapter 3 for a completely randomized trial. However, it would not be obvious whether or not direct adjustment and the Mantel-Haenszel test had fixed the problem. Adjustments might fix the problem—after all, they do fix a problem that looks visibly similar and is produced by randomization on the basis of a covariate. Still, an observational study is not randomized on the basis of a covariate, so one cannot presume that adjustments fix the problem—there is no basis for that presumption. An enthusiastic, perhaps incautious, investigator who declared that adjustments fixed the problem might turn out to be wrong.

The typical criticism of an observational study says, You keep saying your $2 \times 2 \times 4$ table is analogous to a table like Table 5.1, produced by random-

izing on the basis of a covariate, but actually there is another covariate, an unobserved covariate u_i (say, education recorded in four levels), and had you disaggregated your $2 \times 2 \times 4$ table into a $2 \times 2 \times 16$ table using u_i (separating people by education), the results would have been completely different. Your stratum 1 for older men, the critic continues, is actually analogous to Table 5.2: it is a marginal table collapsed over education, over u_i, and the argument that showed Table 5.2 to be misleading is precisely the criticism I am making of your adjustments; you adjusted for x_i, but you needed to adjust for both x_i and u_i. The incautious investigator and the critic are both appealing to the same facts about a marginal table collapsed over a variable that predicts both treatment assignment and mortality.

Crying Babies

Why Do Babies Cry?

Why do babies cry? A baby may cry because it is hungry or because it is uncomfortable. Sometimes babies are just irrational and helplessly cry for no reason, and it takes an in-control, rational adult to calm the baby by rocking it. At least, this is the story that parents tell each other. The story that babies tell each other is a little different, as demonstrated in a curious randomized experiment by T. Gordon and B. M. Foss.[11]

The experiment in Table 5.1 is a simulation, and I have been saying that it is peculiar to randomly assign treatments with different probabilities in different strata. The Gordon and Foss experiment has this peculiar feature and other interesting features as well.

Rocking a Random Baby

Gordon and Foss hypothesized that a baby may cry without being uncomfortable or hungry for food, because instead it is hungry for stimulus. That is, a baby may cry because it wants to be rocked. Perhaps the baby knows what it wants, is active in getting what it wants, and knows how to get what it wants. Gordon and Foss write, "Since rocking seems an effective means of stopping the baby from crying when it is not hungry or in pain, it would

appear to answer some of the stimulus needs of the young infant . . . That is to say, if crying can result from stimulus hunger then appropriate stimulation should put off the onset of crying."[12]

The experiment involved full-term babies in a ward of an obstetrics hospital. After their noon feeding, the babies spent an hour and a half together in the nursery. On each day for 18 days, the babies in the ward who were not crying were identified, and one was picked at random and rocked for 30 minutes, with the other identified babies serving as controls.

On the first day, there were nine identified babies who initially were not crying. One of the nine was rocked, yielding eight controls. Over the course of the half hour, the one rocked baby did not cry, but five of the eight control babies cried. On the second day, the identified list contained seven babies: one rocked baby and six controls; the rocked baby did not cry, but four of the six control babies did cry. In the end, the rocked babies cried less. This is not inconsistent with Gordon and Foss's hypothesis that the babies were crying because of stimulus hunger, essentially because they wanted to be rocked.

Notice that the chance that any one baby would be rocked varied from day to day because the identified list of babies contained a different number of babies on different days, and only one baby was rocked on each day. On day one, there were nine babies on the identified list, so the chance of being rocked was 1/9 for each baby; but on day two there were seven babies, and the chance of being rocked was 1/7 for each baby. In this sense, the babies experiment closely resembles the simulated experiment in Table 5.1, now with 18 strata, one for each day, and one treated baby on each day. The chance of being rocked varied from day to day, from stratum to stratum, but on any given day, within any given stratum, the chance of being rocked was the same for every baby.

✳   Interference between Units

So far, causal effects have been described as comparisons of two potential responses a person might exhibit: the response r_{Ti} that person i would exhibit if treated and the response r_{Ci} that person i would exhibit if in the control group. In this description, baby i will cry or not if rocked, r_{Ti}, and will cry or not if not rocked, r_{Ci}; so the effect of rocking on baby i is $\delta_i = r_{Ti} - r_{Ci}$. At least potentially, the experiment by Gordon and Foss illustrates one way

in which that description may be inadequate, called "interference in treatment effects between units"—here, interference between babies.

There is no interference between babies if whether baby i cries depends upon whether baby i is rocked but not on whether some other baby is rocked. Generally, there is no interference between experimental units if the response of a unit may depend on the treatment received by that unit but not on the treatments received by other units. There is no interference between people in an experiment if Harry responds to the treatment received by Harry and Sally responds to the treatment received by Sally—that is, if Harry is unchanged by the treatment Sally receives, and Sally is unchanged by the treatment Harry receives. No interference seems plausible, if not inevitable, for the ProCESS Trial in Chapter 1. Is it plausible in the crying babies experiment?

Babies sometimes cry in response to an unpleasant stimulus. One unpleasant stimulus is the sound of a crying baby. Prevent one baby from crying by rocking it, and you may prevent a symphony of crying, so the treatment given to one baby may affect other babies. "Laugh, and the world laughs with you," wrote the poet Ella Wheeler Wilcox; "Weep, and you weep alone." But perhaps she was not thinking about a hospital nursery.

There is interference between babies if the treatment given to one baby may affect the response of some other baby. On the first day, there were nine babies, hence nine ways to assign one baby to rocking and eight babies to control. Baby $i = 1$, Harry, was a small baby given to gentle whimpering. Whether Harry is rocked has no effect on other babies. By contrast, baby $i = 3$, Sally, was a large, vocal, determined baby with capable lungs. Treat Sally by rocking her, thereby preventing her from crying, and quiet reigns in the nursery; treat any other baby instead, and Sally's crying would ensure widespread crying. If there is interference between babies, then on day 1, with nine ways to pick one baby for treatment, every baby has nine potential responses, a different potential response for each choice of a baby to rock. When there is interference between units, causal effects are no longer comparisons of two potential responses unit i might exhibit, $\delta_i = r_{Ti} - r_{Ci}$, because unit i has different potential responses depending on the treatments given to other units.

Fisher's null hypothesis of no effect implies no effect, hence no interference in effects. As a consequence, the Mantel-Haenszel test may be used to test Fisher's hypothesis. As soon as one steps beyond testing Fisher's hypothesis of no effect, interference complicates causal inference.[13]

✳ Repeated Use of the Same Subjects

In Gordon and Foss's experiment, some of the babies in the nursery on day one were still there on day two. A baby might spend a few days in the nursery, hence participate in the experiment several times. Sally was rocked on day one, but she was a control baby on day two. Generally, an experimental unit—a neutral term—is an opportunity to apply or withhold treatment. In Gordon and Foss's experiment, each baby contributed several experimental units, several days on which the baby might be rocked or used as a control. Previously in this book, each person was one experimental unit, but in the crying babies experiment the two concepts diverge: one baby is several experimental units because the baby stayed in the nursery for several days.

Repeated use of the same babies has various consequences. It is an additional opportunity for interference between units. Perhaps rocking Sally on day one affects whether she cries on day two—perhaps she came to the nursery on day two expecting to be rocked again, and she was vocal in expressing her disappointment that this did not occur. When one person is several experimental units, interference between that person's several units is not unlikely.

If we study a few babies many times, we learn more about those particular babies but perhaps less about babies in general. In survey sampling, there is a strategy known as clustered sampling or multistage sampling.[14] For instance, the High School and Beyond longitudinal study found it practical to sample U.S. high school students by first sampling high schools—the clusters or primary sampling units—then sampling students inside the sampled high schools, the units inside the clusters.[15] In survey research, multistage sampling may reduce cost but may increase the imprecision of estimates as a description of the population as a whole. By a weak and not always apt analogy with survey sampling, a person who contributes several experimental units to an experiment is sometimes described as a "cluster of units": several experimental units are "clustered" in that they are observations on the same baby. Clustering is an entirely different issue from interference, although confusion on this score is not uncommon.

If individual units, not clusters, are randomized—if a baby is picked at random on each day—then a randomization test, such as the Mantel-Haenszel test, may be used to test Fisher's hypothesis of no treatment effect.

Repeated use of the same babies becomes relevant when moving beyond a test of no effect.

In the extreme, there is only one person in the experiment, and all of the experimental units refer to this person, who is repeatedly randomized to treatment or control, perhaps on different days. The most famous example of a single-subject randomized experiment is the "lady tasting tea," described by Fisher in chapter 2 of his book *Design of Experiments.* This experiment concerned the claimed ability of a particular lady to distinguish the taste of tea to which milk was added, as opposed to milk to which tea was added. Notice that the hypothesis in this experiment concerns the ability of a particular lady, not the typical ability of people in general.

Single-subject randomized experiments are of increasing practical importance. They are increasingly being used to pick the best treatment for a particular person, acknowledging that the best treatment for one person may not be best for another.[16]

Unlike the crying babies experiment in which each baby was several units, in some other experiments groups or clusters of people form a single unit, and whole clusters are randomly assigned to treatment or control.[17] For example, the PROSPECT (Prevention of Suicide in Primary Care Elderly: Collaborative Trial) randomized trial asked whether adding a specialist nurse—a depression care manager—to a primary care medical practice would improve the treatment of depression at the practice.[18] Twenty primary care practices were randomly divided into two groups of 10 practices, and a depression care manager was added to each of the 10 practices in the first group. So this experiment had only 20 units—the 20 practices that were randomized—even though more than 8,000 patients were screened for depression at these 20 practices during the trial.

Combining Strata with the Same Probability of Treatment

Combining Strata in the Simulated Example

The simulated example in Table 5.1 was constructed with no treatment effect. Table 5.1 correctly gave no sign of a treatment effect, but when Table 5.1 was collapsed to its marginal 2 × 2 table, Table 5.2 gave the misleading impression

Table 5.3.   A collapsed version of Table 5.1, but stopping with two strata defined by
              age (which happens to determine the propensity score in this example)

		Stratum 1: Older folks, probability of treatment = 0.8		
Group	*Dead*	*Alive*	*Total*	*Morality rate (%)*
Treated	55,851	103,756	159,607	35.0
Control	14,149	26,244	40,393	35.0
		Stratum 2: Younger folks, probability of treatment = 0.2		
Group	*Dead*	*Alive*	*Total*	*Morality rate (%)*
Treated	6,014	33,805	39,819	15.1
Control	23,986	136,195	160,181	15.0

that the treatment was harmful, killing perhaps 12% = 31% − 19% of people exposed to it. Each of the four strata of Table 5.1 is a completely randomized experiment, but if the strata are ignored, then the result in Table 5.2 is far from a completely randomized experiment.

Consider Table 5.3, a table between Tables 5.1 and 5.2. Table 5.3 collapses Table 5.1 over gender, but not over age. Stratum 1 of Table 5.3 consists of older folks, whether men or women, while stratum 2 consists of younger folks, whether men or women. For instance, in the upper left corner of Table 5.3, the number of deaths among older treated individuals is 55,851, and this is obtained from Table 5.1 by adding the corresponding entries for older men and for older women, 55,851 = 31,868 + 23,983. Each entry in Table 5.3 is the sum of two entries from Table 5.1, whereas each entry from Table 5.2 is the sum of four entries from Table 5.1.

Table 5.3 is simpler than Table 5.1 in having two strata rather than four, but like Table 5.1 it leaves the correct impression that the treatment had no effect, unlike Table 5.2. If we test Fisher's null hypothesis of no effect in stratum 1 of Table 5.3, the two-sided *P*-value is 0.90, whereas in stratum 2 the *P*-value is 0.52. Most randomized treatment assignments would produce mortality rates that differ more than the rates in stratum 1 of Table 5.3, namely, 35.0% and 35.0%, and more than the rates in stratum 2 of Table 5.3, namely, 15.1% and 15.0%. If the two strata are used in a single randomization test, the Mantel-Haenszel test, the single test of no treatment effect yields a two-sized *P*-value of 0.78, so again there is no evidence suggesting a treatment effect. In this sense, Table 5.3 is not misleading.

There is a second sense in which Table 5.3 is not misleading. If direct adjustment is applied to Table 5.3 using stratum total weights, the average treatment effect is correctly estimated to be close to zero. Stratum 1 and stratum 2 of Table 5.3 each contain 200,000 people, so weights that are proportional to the stratum size will give equal weights to the two strata: $200{,}000 / 400{,}000 = 1/2$. Unlike Table 5.2, direct adjustment will correct for the fact that most people in stratum 1 of Table 5.3 received treatment, and most people in stratum 2 received control. The directly adjusted estimate of the average treatment effect based on Table 5.3 is

$$(35.0 - 35.0) / 2 + (15.1 - 15.0) / 2 = 0.005,$$

or five one thousandths of 1%, virtually zero.

Although Table 5.3 is not misleading in the two specific senses just mentioned, it does contain less information than Table 5.1. Table 5.1 shows that men are at greater risk than women of the same age, but this is not evident from Table 5.3. There are many important ways in which Tables 5.1 and 5.3 differ, but they agree in the specific sense that neither misleadingly suggests that there is a treatment effect.

Why is Table 5.2 misleading, when neither Table 5.1 nor Table 5.3 is misleading? Table 5.1 is not misleading because each of its four strata is a completely randomized experiment on its own; that is, the treatment assignment probabilities, $\pi_i = \Pr(Z_i = 1)$, are constant within a stratum even though they do vary from one stratum to another. No trouble occurs in Table 5.1, provided we do not merge its separate experiments. Table 5.3 did merge the experiments in Table 5.1, but in a special way. In the simulation that produced Table 5.1, treatments were assigned by independent coin flips, in which older men and women received treatment with probability $0.8 = \pi_i = \Pr(Z_i = 1)$ and younger men and women received treatment with probability $0.2 = \pi_i = \Pr(Z_i = 1)$. Stratum 1 of Table 5.3 merged the strata in Table 5.1 that had a $0.8 = \pi_i = \Pr(Z_i = 1)$ chance of treatment, whereas stratum 2 of Table 5.3 merged the strata in Table 5.1 that had a $0.2 = \pi_i = \Pr(Z_i = 1)$ chance of treatment. That is, Table 5.3 merged strata that had the same probability of treatment, $\pi_i = \Pr(Z_i = 1)$, and avoided merging strata with different probabilities of treatment. The two strata in Table 5.3, like the four strata in Table 5.1, are each completely randomized experiments: everyone in the same stratum had the same probability of treatment.

Table 5.4. A collapsed version of Table 5.1, but stopping with two strata defined
by gender

		Stratum 1: Men		
Group	*Dead*	*Alive*	*Total*	*Morality rate (%)*
Treated	35,861	63,988	99,849	35.9
Control	24,139	76,012	100,151	24.1
		Stratum 2: Women		
Group	*Dead*	*Alive*	*Total*	*Morality rate (%)*
Treated	26,004	73,573	99,577	26.1
Control	13,996	86,427	100,423	13.9

Table 5.4 collapses Table 5.1 over age rather than over gender, and now the
problem seen in Table 5.2 has recurred: the treatment incorrectly appears to
harm men and to harm women. The problem recurs because two randomized
experiments with different probabilities of treatment π_i are merged.

Collapsing a Table and Covariate Balance

Tables 5.1 to 5.4 illustrate another fact concerning covariate balance. In
Table 5.3, gender is not explicitly present, but it is balanced in the treated
and control groups within each age stratum, whereas in Table 5.4 age is sub-
stantially out of balance in the treated and control groups. By examining
Table 5.1, we see, for instance, that in Table 5.3 in the first stratum, among
older folks, the treated group is 50.0% = 79,779 / (79,779 + 79,828) female
and the control group is 50.0% = 20,221 / (20,221 + 20,172) female. That is,
although Table 5.3 collapsed over gender, in each stratum of Table 5.3 the
treated and control groups are comparable in terms of gender, so it is not
unreasonable to compare the treated and control groups. The situation is very
different in Table 5.4. In the first stratum of Table 5.4, among men the treated
group is 79.9% = 79828 / (79828 + 20021) older men, while the control group
is 20.1% = 20172 / (20172 + 79979) older men. That is, in Table 5.4 it is not
reasonable to compare the treated and control men because the treated and
control groups differ so much in age.

Expressed concisely, in the peculiar randomized trial, collapsing strata
with the same treatment assignment probabilities, $\pi_i = \Pr(Z_i = 1)$, preserved

key features of the randomized design and balanced the collapsed covariates. In contrast, collapsing strata with different treatment assignment probabilities, $\pi_i = \Pr(Z_i = 1)$, destroyed key features of the randomized design and left the collapsed covariates imbalanced.

Matched Pairs

Matched Pairs of One Treated Person and One Control

A simple study design has $P \leq I/2$ pairs in which each pair has one treated individual and one control. For instance, if the first pair contained the first two people, say, Harry with $i = 1$ and Sally with $i = 2$, then either Harry would be treated and Sally would be a control with $Z_1 = 1$ and $Z_2 = 0$, or else Harry would be a control and Sally would be treated with $Z_1 = 0$ and $Z_2 = 1$, so in either case $Z_1 + Z_2 = 1$. If I tell you that $Z_1 + Z_2 = 1$, then I have told you that either Harry or Sally received treatment and the other received control, but I have, so far, declined to tell you whether it was Harry who received treatment or whether it was Sally.

Imagine that in a population before people are matched the investigator assigns people to treatment, $Z_i = 1$, or control, $Z_i = 0$, by unrelated flips of coins with varying probabilities of a head, $\pi_i = \Pr(Z_i = 1)$. The investigator then pairs people for whom the same coin was used, with the added requirement that one is treated and one is control. Each person is in at most one pair: Harry appears in at most one pair, and so does Sally. A curious and useful thing happens if you do this.

For example, the investigator used the $1/3$-coin for both Harry with $i = 1$ and Sally with $i = 2$, so $\pi_1 = \pi_2 = 1/3$. By luck of the coin flips, exactly one of Harry and Sally was actually assigned to treatment, so $Z_1 + Z_2 = 1$. In this case, the investigator could pair Harry and Sally because they had the same π_i, and exactly one was treated, that is, $\pi_1 = \pi_2$ and $Z_1 + Z_2 = 1$. Other pairs are formed by pairing two people with $\pi_i = \pi_j = 1/4$ and $Z_i + Z_j = 1$, and so on. That is, each pair has $\pi_i = \pi_j$ and $Z_i + Z_j = 1$, but the common value of $\pi_i = \pi_j$ changes from one pair to the next pair, perhaps $1/3$ in one pair, $1/4$ in another, $1/2$ in still another.

Suppose that the investigator did just that, namely, noticed that (i) $\pi_1 = \pi_2 = 1/3$ and (ii) $Z_1 + Z_2 = 1$, and on the basis of (i) and (ii) alone decided to pair Harry and Sally. Now, knowing this, I have a question for you. Who is

more likely, Harry or Sally, to have received treatment in this matched pair? Intuitively, the situation is perfectly symmetrical, so that, having matched in this way, Harry and Sally are equally likely to be the one treated person in this pair.[19] In fact, this intuition is correct, and within a pair constructed in this way it is a fair coin flip that Harry rather than Sally was treated; that is, the probability is 1/2. Notice that in this argument it was critically important that $\pi_1 = \pi_2$ and completely unimportant that the common value was $1/3 = \pi_1 = \pi_2$. No matter what the common value of $\pi_1 = \pi_2$ is, the same argument shows that pairing with $\pi_1 = \pi_2$ and $Z_1 + Z_2 = 1$ means that the two people now have an equal chance, $1/2$, of being the treated person in the matched pair.

In this simple design, the unequal probabilities $\pi_i = \Pr(Z_i = 1)$ have disappeared, and within each pair each of the two paired people has probability 1/2 of being the treated person rather than the control. In other words, this simple design has removed unequal probabilities of treatment and recovered one of the simplest randomized experiments, namely, P pairs of two people, with one person in each pair picked at random for treatment by unrelated flips of a fair coin, the other person receiving control.

So what has happened? We saw earlier in this chapter that if you randomize with unequal probabilities $\pi_i = \Pr(Z_i = 1)$ but put people with the same probability of treatment into the same stratum, then with some extra work in analysis you can patch things up. For example, using direct adjustment you can estimate average treatment effects, and using the Mantel-Haenszel test you can test Fisher's hypothesis of no effect. In contrast, the matched design has patched things up in the design of the study, and by matching has recovered a simple randomized experiment. Again, all of this occurs in a fictional world that lives between randomized experiments and observational studies, a world in which treatments are assigned to subjects with unequal probabilities, $\pi_i = \Pr(Z_i = 1)$, to which the investigator has access, so the investigator can pair people who share the same probability of treatment, $\pi_i = \Pr(Z_i = 1)$.

Matched Pairs under Fisher's Hypothesis of No Treatment Effect

Suppose that we have paired Harry, $i = 1$, and Sally, $i = 2$, knowing that exactly one was treated, $Z_1 + Z_2 = 1$, and let us write Y for the treated-

minus-control response in this pair. If Harry receives treatment in this pair, then Y is Harry's response minus Sally's response, namely, $R_1 - R_2$; but if Sally received treatment, then Y is Sally's response minus Harry's response, namely, $R_2 - R_1$.

Suppose Fisher's hypothesis H_0 of no treatment effect is true, that is, suppose $r_{Ti} = r_{Ci}$, for every person i. This means that we observe r_{C1} from Harry, person $i = 1$, whether Harry receives treatment with $Z_1 = 1$ or Harry receives control with $Z_1 = 0$. In the same way, we observe r_{C2} from Sally, person $i = 2$, whether Sally receives treatment with $Z_2 = 1$ or Sally receives control with $Z_2 = 0$. In either case, $Y = \pm(r_{C1} - r_{C2})$. Moreover, under the same conditions, the absolute value of the difference, $|Y| = |r_{C1} - r_{C2}|$, does not change when the treatment assignments, Z_1 and Z_2, change.

In addition to supposing Fisher's hypothesis H_0, suppose that we have paired Harry and Sally because $\pi_1 = \pi_2$ and $Z_1 + Z_2 = 1$. Then within the pair, with probability 1/2, Harry received treatment in this pair, and with probability 1/2 it was Sally who received treatment. Therefore, $Y = \pm(r_{C1} - r_{C2})$, and the positive and negative values each occur with probability 1/2.

In general, if persons i and j are paired to form pair p because $\pi_i = \pi_j$ and $Z_i + Z_j = 1$, and if there is no treatment effect, then the treated-minus-control pair difference Y_p in this pair is $Y_p = \pm(r_{Ci} - r_{Cj})$, with the positive and negative values each occurring with probability 1/2. In other words, we know what "no treatment effect" would look like in this setting: we would see treated-minus-control pair differences Y_p whose histogram or distribution would tend to be symmetric about zero. That is, we would see $Y_p = 1$ and $Y_p = -1$ with equal probability, $Y_p = 2$ and $Y_p = -2$ with equal probability, and so on. An asymmetric distribution of pair differences, Y_p—say with more $Y_p = 2$ than $Y_p = -2$—is either a sign that the treatment had some effect or a sign of some bias in treatment assignment, $\pi_i \neq \pi_j$.

Matched Pairs with Binary Outcomes

The case of binary outcomes is particularly simple because then r_{Ci} is either 1 or 0, and $\pm(r_{Ci} - r_{Cj})$ is one of $0 = 1-1$, or $0 = 0-0$, or $1 = 1-0$, or $-1 = 0 - 1$. A pair p is said to be concordant if the treated-minus-control pair difference Y_p is 0; otherwise, if Y_p is ± 1 it is said to be discordant. Moreover, finding more $Y_p = 1$ than $Y_p = -1$ is either a sign of a treatment effect or a sign of a bias in treatment assignment. Because the pairs can produce only four possible

results, it is convenient to create a table of the frequencies of these four possible results. Such a table counts pairs, not people, and it records the binary outcome for the treated person, 1 or 0, in the rows, and the binary outcome for the control, again 1 or 0, in the columns. The table's diagonal contains the concordant pairs; the off-diagonal contains the discordant pairs.

Let us consider an example.

Delirium in the Hospital

Delirium is defined as a state of acute confusion. It is common among the elderly when hospitalized. Sharon Inouye and colleagues conducted a nonrandomized, matched evaluation of an intervention designed to reduce the frequency of delirium among the elderly when hospitalized.[20] The intervention or treatment, called the Elder Life Program, was administered by a team including a nurse specializing in geriatric nursing, a physical therapist, and others. The protocol had numerous elements, including ensuring that elderly patients received needed vision and hearing aids, daily conversation and cognitive activity, daily physical activity, quiet and a warm drink at bedtime, and early recognition of dehydration.

A randomized trial was judged impractical, so instead the study intervened at one unit of the hospital, and then matched patients from that unit with similar patients treated at other units where there was no intervention and the patients were treated in the usual way.[21] The matching paired patients for age, gender, and several predictors of delirium, including vision impairment, severe illness, and cognitive impairment. So, in the end, there were pairs of similar elderly patients, one treated, the other control, so if patients i and j are paired to form pair p then $Z_i + Z_j = 1$. Presumably, the investigators hoped to pair so that $\pi_i = \pi_j$—that is, so that only luck distinguished the two paired people before the intervention—but the investigators could not know this to be true.

Inouye and colleagues began their analysis with a table analogous to Table 1.1 showing that the intervention and matched control groups looked similar upon admission to the hospital, before the start of the intervention. For instance, the groups were similar in terms of age, education, and APACHE II scores, as well as various risk factors for delirium. Unlike Table 1.1, which came from a randomized experiment, such a table in a nonrandomized study

Table 5.5. Delirium in 852 matched pairs of elderly hospitalized patients

		Matched patient receiving control		
	Delirium	Yes	No	Total
Patient receiving treatment	Yes	9	33	42
	No	55	755	810
	Total	64	788	852

is understood simply as describing the covariates in the table, without promises about unmeasured covariates. Inouye and colleagues then looked at the primary outcome, the occurrence or not of delirium, as recorded for matched pairs in Table 5.5.

Table 5.5 counts 852 pairs of two people. There were nine pairs in which both the treated patient and the matched control experienced delirium, and 755 pairs in which neither patient experienced delirium. These $764 = 755 + 9$ pairs are concordant for delirium. There are 55 discordant pairs in which the control experienced delirium but the treated patient did not, and there are 33 discordant pairs in which the treated patient experienced delirium but the control did not.

About 5% of the treated group ($42/852$) and 7.5% of the control group ($64/852$) experienced delirium. The evidence relevant to Fisher's hypothesis of no treatment effect comes from the discordant pairs: their number, $88 = 55 + 33$, and the split, 55 versus 33. If Fisher's hypothesis H_0 of no treatment effect were true, then the concordant pairs would exhibit the same pattern of delirium had the treatments assignments been reversed. That is, if H_0 is true and if pair p consists of patients i and j, and pair p is concordant with $0 = Y_p = \pm(r_{Ci} - r_{Cj})$, then Y_p would still equal zero if the treatment assignment were reversed in this pair. If there is no effect in a discordant pair, then reversing the treatment assignment would swap the roles of the two patients without changing their delirium, and the treated minus control difference would change sign. That is, if H_0 is true and if pair p consists of patients i and j, and pair p is discordant with $1 = |Y_p| = |r_{Ci} - r_{Cj}|$, then reversing the treatment assignment in pair p would replace Y_p by $-Y_p$.

The 55 versus 33 split of discordant pairs in Table 5.5 represents an excess of delirium in the control group. If Fisher's hypothesis H_0 of no effect were true, and if there were no bias in treatment assignment so $\pi_i = \pi_j$ whenever

i and j are paired, then the split into 55 versus 33 would have to have occurred from $88 = 55 + 33$ fair coin flips. In fact, 88 fair coin flips will produce 33 or fewer heads with probability 0.0123 so that the two-sided P-value is $2 \times 0.0123 = 0.0246$.[22] The 55 versus 33 split, or a more imbalanced split, would be a fairly rare event—an event with probability 0.0246—if there were neither a treatment effect nor a bias in treatment assignment. Had this been a paired randomized experiment, then biased treatment assignment would not be an explanation, and the 55 versus 33 split would constitute moderately strong evidence that the treatment protocol reduced delirium.

McNemar's Test

The test we just conducted compared the 55 versus 33 split among discordant pairs in Table 5.5 to $88 = 55 + 33$ fair coin flips. This test is known as McNemar's test, and it is the randomization test of Fisher's hypothesis H_0 of no effect in a paired randomized experiment with binary outcomes.[23] McNemar's test is a special case of the Mantel-Haenszel test in which the P pairs become P strata in a $2 \times 2 \times P$ contingency table, but the layout in Table 5.5 is far more compact.[24]

The Propensity Score

What Is the Propensity Score?

Pick a value x of the observed covariates, any one value you prefer. In Table 5.1, there are four possible values of x: for older men, older women, younger men, or younger women. For each value x of the observed covariates, the propensity score is a number, λ_x: it is the average value of the treatment assignment probabilities $\pi_i = \Pr(Z_i = 1)$ over all individuals i whose observed covariate x_i is x[25]. If the x that you picked was for younger women, then in Table 5.1 $\pi_i = \Pr(Z_i = 1) = 0.2$ for every younger woman, and the propensity score for younger women, λ_x, is the average of the 100,000 π_i values for younger women, each of which is $\pi_i = 0.2$, so their average is $\lambda_x = 0.2$. The propensity score is extremely simple in an experiment that is randomized on the basis of a co-

variate like Table 5.1 because, by the definition of randomization on the basis of a covariate, people with the same x_i have the same π_i.

In an observational study, not an experiment randomized on the basis of a covariate like Table 5.1, the propensity score, λ_x, is an average of treatment assignment probabilities π_i that may vary among people with the same value x of the observed covariate x_i. In an observational study with the same observed covariates as in Table 5.1, two younger women—say, women i and j—would have $x_i = x_j$ because they are both younger women, but they might have $\pi_i \neq \pi_j$ because they differ in some other way that affects their treatment assignment probabilities, perhaps because they differ in terms of an unmeasured covariate, $u_i \neq u_j$. This is one sense in which an experiment randomized on the basis of a covariate is quite an oversimplification of what happens in an observational study. In an observational study, people who look similar before treatment may, and often do, differ in ways that distort inferences about treatment effects.

Propensity Scores Balance Observed Covariates *x*, Not Unobserved Covariates *u*

We saw that collapsing Table 5.1 over gender in Table 5.3 produced strata in which gender was balanced. This occurred because we collapsed strata with the same treatment assignment probabilities, $\pi_i = \Pr(Z_i = 1)$. What would happen if we collapsed strata with the same propensity score, λ_x, but perhaps different treatment assignment probabilities?

Table 5.1 is from a simulation, so we know that $\pi_i = 0.8$ for older folks and $\pi_i = 0.2$ for younger folks, but if Table 5.1 had come from an observational study then we would not know this. We can see from Table 5.1 that a reasonable estimate of the propensity score, λ_x, for older men is $79,828 / 100,000 = 0.798$, close to 0.8, and a reasonable estimate of the propensity score, λ_x, for younger women is $19,798 / 100,000 = 0.198$, close to 0.2. That is, from the data we can see that the average probability of treatment, λ_x, is close to 0.8 for older men, but we cannot know whether each and every older man had a 0.8 probability of treatment, $\pi_i = 0.8$. Clearly, what we can see, by itself, is not enough to justify testing the hypothesis of no treatment effect for older men by applying the argument from Chapter 3

to the first stratum of Table 5.1; to justify that, we need a completely randomized experiment for older men, and that means $\pi_i = 0.8$ for every older man. Would knowing λ_x be of any value on its own?

We saw that when we collapsed Table 5.1 to form Table 5.3, gender was not explicitly mentioned in Table 5.3 but it was balanced: within each of the two strata of Table 5.3 the treated and control groups had similar proportions of women. It turns out that this balancing property does not require knowledge of π_i for each of the $i = 1, \ldots , I = 400{,}000$ people in Table 5.1, just knowledge of λ_x for the four strata or values of x in Table 5.1.

The balancing property of propensity scores states that if you collapse strata with different x but the same λ_x, then you balance the distribution of the parts of x that are no longer explicitly mentioned in the collapsed table. It is important to understand the content and limitations of this property, and the remainder of this section is devoted to clarifying the content and limitations. The property is not difficult to prove, essentially using Bayes' theorem.[26]

First, this property refers to balancing covariates in the sense that the observed covariates in Table 1.1 were balanced in the ProCESS Trial: the distributions of x in the treated and control groups in Table 1.1 are not identical, but they reflect only chance imbalances, the type produced by fair coin flips. In Tables 5.1 and 5.3, the sample size is so large that chance imbalances in proportions and averages are very small. Whether the sample size is large or small, the balancing property of propensity scores says that the only imbalances in the collapsed part of x are chance imbalances.

A key limitation of the balancing property of propensity scores is that it falls far short of the balancing property of randomized treatment assignment. Randomization in the ProCESS Trial balanced both observed covariates x and unobserved covariates u because a fair coin was flipped to assign treatments, and it produced only the chance imbalances that fair coins produce. In contrast, propensity scores balance observed covariates x but not unobserved covariates u. Unlike fair coin flips, which do not use the covariates, collapsing values of x with the same propensity score λ_x uses a specific list of observed covariates x. To use propensity scores to balance both x and an unobserved covariate u, you would need $\lambda_{(x,u)}$, the propensity score for x and u together, where $\lambda_{(x,u)}$ is the average π_i for all individuals i with $x_i = x$ and $u_i = u$; however, you have no way to determine or estimate $\lambda_{(x,u)}$ because you have not observed u_i. You can only use propensity scores to balance observed co-

variates x because the activity is a process of collapsing over values of x with the same λ_x.

We would not know the value of λ_x in an observational study because that entails knowing the π_i, and we know the π_i only if we randomly assign treatments. However, as we have seen in the case of Table 5.1, we can use the data to estimate λ_x in an observational study. Because the sample size in Table 5.1 is so large, the estimates $\hat{\lambda}_x$ of λ_x are close to their true values. How much harm is done by having to use estimates of λ_x rather than true values?

Usually estimates work less well than true parameters. In a randomized experiment, an estimate depends upon the coin flips that assigned treatments whereas a population parameter does not. In Chapter 2, the population parameter $\bar{\delta} = \bar{r}_T - \bar{r}_C$, the average treatment effect, was estimated by $\hat{r}_T - \hat{r}_C$ in Table 3.1, but this estimate was affected by the random division of the ProCESS patients into treated and control groups, and we would much prefer to have the population parameter, $\bar{\delta}$. Similarly, in Chapter 3 we used the estimates in Table 3.1 to test the null hypothesis of no treatment effect, $\delta_i = 0$ for all i, but life would have been easier if someone had simply handed us all of the δ_i values. Usually, estimates work less well than true parameters.

Although typically correct, the statements in the previous paragraph are too vague to be serviceable. In particular, "work less well" is vague because whether something works well depends upon what work you want done. If your goal were to estimate the propensity score, then you would prefer the true value λ_x to the estimate $\hat{\lambda}_x$, but that is not the goal. When propensity scores are used, the goal is to draw inferences about causal effects adjusting for observed covariates x. In collapsing values of an observed covariate, the goal is to balance the parts of x that are not given explicit consideration. It turns out that, for this goal, estimates $\hat{\lambda}_x$ are slightly better than true propensity scores λ_x. Why is that? Sample estimates err by being close to the sample, rather than close to the population. Sample estimates of a propensity score, $\hat{\lambda}_x$, cannot tell the difference between an imbalance in x produced by unequal treatment assignment probabilities and an imbalance in x produced by unlucky coin flips. These imbalances look the same in data, and $\hat{\lambda}_x$ is based on data. So the sample estimates of $\hat{\lambda}_x$ tend to correct both imbalances in x that occur systematically and those that occur by chance, making the balance on observed covariates a tad better than expected by chance alone.[27] In this rather specific sense, estimated propensity scores, $\hat{\lambda}_x$, can "work better" than true propensity scores, λ_x. Alas, this is only true as

a statement about observed covariates; propensity scores, true or estimated, do little or nothing to address imbalances in unobserved covariates.

Using Propensity Scores to Form Balanced Strata

Propensity scores are not used to collapse four strata to two strata, as in the step from Tables 5.1 to Table 5.3. When used with strata, propensity scores produce a small number of strata that balance a large number of covariates.

If there are C covariates, each at L levels, then how many strata are there? In Table 5.1, there are $C = 2$ covariates, age and gender, each at $L = 2$ levels, hence $4 = 2^2$ strata. In general, C covariates each at L levels yields L^C strata. Table 5.6 shows how the number of strata changes with the number of covariates when each covariate has either $L = 2$ or $L = 3$ levels.

The important feature of Table 5.6 is the explosion in the number of strata as the number of covariates increases. Think about it this way. As I write, there are roughly 7.5 billion people on Earth. With a mere $C = 30$ covariates, each at $L = 3$ levels, there are about $3^{30} = 2.1 \times 10^{14}$ strata, or more than 27,000 strata for every person on Earth. Even if a study included everyone on Earth, most strata would be empty. Unless a few strata were extremely popular, it is likely that many if not most occupied strata would contain a single person, hence providing no comparison of similar people receiving treatment and control. The explosion in Table 5.6 of L^C as C increases shows that the desire to compare identical people is a forlorn desire—it cannot be done if we recognize more than a few ways people can differ.

Table 5.6. How many strata are there with C covariates each at L levels?

Number of covariates (C)	L = 2 levels	L = 3 levels
1	2	3
2	4	9
5	32	243
10	1,024	59,049
20	~1 million	~3.5 billion
30	~1 billion	~2.1×10^{14}
50	~1.1×10^{15}	~7.2×10^{23}
75	~3.8×10^{22}	~6.1×10^{35}

In contrast, the first application of propensity scores involved $C=74$ covariates in a study of the survival and functional status of 1,515 patients with coronary artery disease, of whom 590 were treated with coronary artery bypass graft surgery and the remaining 925 were treated with drugs.[28] An estimate, $\hat{\lambda}_x$, of the propensity score, λ_x, was obtained using the covariates in a model designed to fit the observed treatment assignments—that is, to predict from x_i which patients would receive bypass surgery, $Z_i=1$.[29] Unsurprisingly to cardiologists, the model judged the patients most likely to receive bypass surgery to be experiencing chronic chest pain, to have narrowing of the left main coronary artery, and to have reasonably intact left ventricular function. Patients were grouped into five strata using the estimates, $\hat{\lambda}_x$, one more stratum than in Table 5.1, each stratum containing $303=1{,}515/5$ or 20% of the patients with similar $\hat{\lambda}_x$. The covariate balance was checked, as it always should be checked, by comparing the distribution of x for treated and control subjects in the same stratum. Within these five strata, all $C=74$ covariates were in balance, indeed in slightly better balance than expected by random assignment of treatments. The better-than-expected balance results from the use of estimates, $\hat{\lambda}_x$, in place of true values, λ_x, as discussed in the previous section. Unlike randomization, balancing the observed covariates x using propensity scores does little or nothing to control imbalances in unobserved covariates u.

Contrasting this example with Table 5.6, we see that it is virtually impossible to find people who are identical in terms of $C=74$ covariates, but it is comparatively straightforward to balance $C=74$ observed covariates.

Strata formed from propensity scores often balance covariates, but they may blur some distinctions that we prefer to keep intact. Collapsing Table 5.1 to Table 5.3 balanced gender, but it blurred the higher risk of death faced by men. In the bypass surgery example, the stratum of patients least likely to receive bypass surgery included some patients who were too sick for surgery, with very poor left ventricular function, and others who were too healthy for surgery, with limited coronary stenosis. It would aid interpretation to keep the sickest and healthiest patients apart, even if that is not needed to balance the observed covariates. Propensity scores are the coarsest summary of the observed covariates x that balance the observed covariates.[30] However, a finer stratification, perhaps a stratification using λ_x and a few key observed covariates, would also balance x. For instance, in the bypass surgery example, comparisons of functional improvement used 45 strata rather than

5 propensity score strata, where the 45 strata were formed from 5 propensity strata, by 3 strata defined by the number of diseased coronary arteries, by 3 strata giving pretreatment functional status, making $45 = 5 \times 3 \times 3$.

Advantages of Matching over Stratification

When, as is often true, we prefer to balance covariates and preserve additional distinctions, a better strategy than stratification is to match for propensity scores, taking additional steps to control the most important observed covariates.[31] In the simplest case, pairs are formed, with one treated person paired with one control as in the study of delirium in elderly patients. When many controls are available, each treated person may be matched to several controls, in parallel with the crying babies experiment where each day had one rocked baby and several control babies. The propensity score permits balancing of many covariates, whereas Table 5.6 indicates that it is not possible to make strata homogeneous in many covariates except by producing many strata that are uninformative in that they contain only one person with no one else for comparison. Matching tests the limits of what is usefully possible by insisting that a matched set contain both a treated subject and a control who are similar in terms of propensity scores and as close as possible in terms of key covariates.[32] A variety of modern devices and algorithms further enhance the advantage of matching over stratification. Matching will appear in many examples in later chapters, and specific matching techniques will be discussed in Chapter 11.

When Is It Enough to Adjust for the Observed Covariates?

Strata are said to be homogeneous in the propensity score, λ_x, if everyone in the same stratum has the same value of λ_x. Strata homogeneous in the propensity score, λ_x, suffice to balance the observed covariates, x, in the sense that only chance imbalances in x remain. As we have seen, strata homogenous in the propensity score do not generally suffice for causal inference, because the probability that patient i is assigned to treatment rather than control,

$\pi_i = \Pr(Z_i = 1)$, may vary not only with observed covariates x_i but also with unobserved covariates u_i. Strata that exhibit balance for observed covariates x_i may be imbalanced for unobserved covariates u_i. People who look the same in measured covariates may, behind the scenes, be very different in ways that were not measured.

It would suffice to adjust for observed covariates, x, if whenever two people, say, person i and person j, had the same observed covariates, $x_i = x_j$, they also had the same probability of treatment, $\pi_i = \pi_j$. This condition, together with the requirement that everyone has a chance of receiving both treatments, $0 < \pi_i < 1$, is called ignorable treatment assignment.[33] When needed for clarity or emphasis, the phrase "ignorable given x" is sometimes used.[34] If treatment assignment were ignorable, then an observational study would be, for all intents and purposes, indistinguishable from an experiment randomized on the basis of a covariate x. As the comparison of Tables 5.1 and 5.2 demonstrated, the analysis of an experiment randomized on the basis of a covariate x needs to take account of x, perhaps by stratifying on x, but if this is done, causal inference is comparatively straightforward.

It is not difficult to show that if treatment assignment is ignorable given covariates x then it is also ignorable given the propensity score λ_x alone.[35] If it suffices to adjust for x then it suffices to adjust for λ_x alone. In the coronary bypass example, if it were sufficient to adjust for the 74 observed covariates recorded in x, then it would be sufficient to adjust for the one variable λ_x. Indeed, if it suffices to adjust for x, then it suffices to adjust for λ_x and any additional aspects of x; see, for instance, the 45 strata in the coronary bypass example in the previous section built from the propensity score and two of the 74 covariates.

In the simulated experiment in Tables 5.1 through 5.4, the treatment assignment was ignorable because x included age. Had x included age but not gender, then the treatment assignment would still be ignorable because people with the same age had the same probability of treatment, $\pi_i = \Pr(Z_i = 1)$. Had x included gender but not age, then the treatment assignment would not have been ignorable because two people with the same gender could have very different probabilities of treatment, $\pi_i = \Pr(Z_i = 1)$, and we would see only the misleading Table 5.4.

The central worry in an observational study is that treatment assignment is not ignorable with observed covariates x because the observed covariates

omit an important unmeasured covariate u, and treatment assignment probabilities $\pi_i = \Pr(Z_i = 1)$ vary with both x and u. Stating this a little more precisely: we worry that treatment assignment would have been ignorable had we measured (x, u), but is not ignorable because we measured x but not u. This would have been the situation in Tables 5.1 to 5.4 had we measured $x =$ gender but failed to measure $u =$ age; then adjustments for (x, u) would have sufficed for causal inference, but adjustments for $x =$ gender alone in Table 5.4 would have been misleading. Stating a worry precisely does not make that worry less worrisome. Stating a concern precisely is the first step in developing plans to address that concern, as we will do in later chapters.

Taking Stock

This chapter has considered a situation that is more complex than a completely randomized experiment but considerably less complex than an observational study. In this chapter, the probability of treatment, $\pi_i = \Pr(Z_i = 1)$, varied with observed covariates, x_i, but two people, say, persons i and j, with the same value of the observed covariates, $x_i = x_j$, had the same probability of treatment, $\pi_i = \pi_j$, a condition called ignorable treatment assignment. For instance, this would occur if we randomly assigned treatments with probabilities that varied with observed covariates x. When treatment assignment is ignorable, an analysis that does not take account of x can yield very misleading estimates of the effects caused treatments, but a straightforward analysis that appropriately adjusts for x removes the problem.

The central problem in an observational study is that an investigator is rarely if ever in a position to claim that treatment assignment is ignorable given the covariates x that happened to have been measured. The investigator is rarely if ever in a position to claim that adjustments for the observed covariates x suffice to estimate treatment effects. The typical criticism of an observational study claims that adjustments for the observed covariate x are not sufficient, that the investigator needed to adjust for both the observed covariate x and a covariate u that was not measured. Scientists should have no fear of forgetting about this central problem because if they do forget, then the referees of their submitted manuscript will remind them about it. And later the author of the editorial discussing their published article will

remind them about it. And the other scientists who publish letters to the editor objecting to their article will remind them about it. And when their article is mentioned in a survey of the status of knowledge on a particular scientific topic, the author of that survey will remind them about it.

Much of the rest of the book speaks to this central problem.

Natural Experiments

Examples of Natural Experiments

In a natural experiment, some key elements of a randomized experiment occur on their own, even though the investigator neither creates nor assigns the treatments. A situation is precarious. A person's fate hangs in the balance. A substantial treatment is given to some people, withheld from others, by a process that is haphazard, senseless, without aim or ambition, equitable, symmetrical.

Lotteries

The obvious examples of natural experiments are lotteries. A lottery is not an experiment, but it is truly randomized.

- What are the effects on work, savings, and consumption of being handed a large pile of cash? Guido Imbens, Donald Rubin, and Bruce Sacerdote examined this question by comparing winners and losers of a state-run lottery in Massachusetts.[1] The median prize for their winners was $635,000. The winners worked less, particularly the winners who were approaching retirement age.

- What is the effect caused by migration from a developing country to a developed one? The Kingdom of Tonga is a chain of islands in the Pacific Ocean, a few hours from New Zealand by plane. In 2002, New Zealand introduced a program called the Pacific Access Category that allowed residents of Tonga to apply to immigrate; then New Zealand held a lottery to permit immigration by 250 winning applicants. Steven Stillman, John Gibson, David McKenzie, and Halahingano Rohorua compared winners and losers of this lottery in terms of subjective and objective measures of well-being.[2] Winners had increased incomes when compared with losers, but the differences in subjective well-being were mixed.

- Does the order of candidates on an election ballot affect the chance of winning the election? During the years from 1978 to 2002, to be fair, California randomized the order of candidates for statewide office. Luck decided which candidate would be listed first or second or last. Exploiting this fact, Daniel Ho and Kosuke Imai looked for an effect of ballot order on election results, concluding that ballot order had no discernable effect on the candidates of major parties, but did affect some candidates from minor parties.[3]

Waiting Lists

Many bureaucratic processes assign the next person in line to the next available treatment. Think of waiting for the next available teller at a post office or a bank where one long queue leads to several tellers. Initially, when you join a long queue, you cannot know to which teller you will be assigned. If there is strict adherence to the waiting list and if one's place on the waiting list is unrelated to one's own attributes—two big ifs—then waiting lists may resemble a lottery. Some observational studies have been built with this thought in mind.

- Does receiving a harsher but reasonable sentence for a criminal conviction reduce recidivism when compared with a more lenient but reasonable sentence? Although this is a perfectly clear causal question, it is not easy to answer because the sentences that convicted felons receive typically reflect consideration of the particulars of the crime and the criminal. It would hardly be surprising if the worst criminals received the

harshest sentences; moreover, if this happened, harsh sentences might appear less effective than they are. Nonetheless, some judges are generally more lenient than others. In the Court of Common Pleas in the state of Pennsylvania, the next available judge in a county is given the next available case. Exploiting this fact within five counties in Pennsylvania, Daniel Nagin and Matthew Snodgrass examined the consequences for recidivism of being sentenced by a harsher rather than a more lenient judge.[4] They were particularly interested in the effects of incarceration, and they concluded, "There is little persuasive evidence that incarceration reduces future criminality."

- If a child grows up in a poorer neighborhood, will that child earn less as an adult? This is a question about the effects of the neighborhood. What would happen to the same child, with the same parents, if the family lived in a different neighborhood? The public housing program in Toronto assigned the next family on the waiting list to the next available residence, sending families to public housing projects in varied neighborhoods of the city, and Philip Oreopoulos exploited this fact to study the effects of neighborhoods on adult earnings.[5]

Studying the Genetic Causes of Disease Using Biological Siblings

Human genetics offers another example of a natural experiment. Except for genes on the sex-linked X and Y chromosomes, Mom and Dad each possess two slightly different copies of each gene, four copies in total, and Mom and Dad each donate one of their two copies to each of their children, who in turn have two copies, one from Mom and one from Dad. Which of Mom's two copies junior receives is essentially random, and the same is true for Dad.[6] So there is a degree of randomization of genes when Mom and Dad have several children.

A common and important study design in human genetics measures the genes of, say, many pairs of two biological siblings, and also measures disease outcomes for these two siblings. Because of the near-random selection of one of Mom's two genes and one of Dad's two genes in forming each sibling, an association between a genetic marker and a disease within pairs of biological siblings is reasonably strong evidence of the presence of some ge-

netic factor in the causation of the disease. Here, the point is not that you and your sister are identical—you are far from identical—but rather that it was simply luck that you received the first copy of one of Mom's two copies of a particular gene and your sister received the second copy. There may be, and often are, subtleties in identifying the precise genetic factor and the biological mechanism through which it produces its effect.[7]

Accidents

There are some interesting and fairly compelling studies in which it is arguable that treatment assignment is close to random, but also arguable that it is not. Accidents are an example. More or less by definition, an accident is not intended or planned; rather, it just happens. Or so we often tell ourselves. In that sense, an accident may seem random. At the same time, some people take precautions to avoid accidents: they wear goggles and are sober when they drive, and they read the safety manual before flicking the switch. Others are reckless: they tailgate while intoxicated with their seat belts unbuckled. Presumably cautious people have fewer accidents—the chip of metal that bounces off goggles passes unremarked and is quickly forgotten. Perhaps cautious people differ from reckless people in many ways.

In an interesting and reasonably compelling study of a difficult topic, Darrin Lehman, Camille Wortman, and Allan Williams asked whether the loss of a spouse or a child in a motor vehicle crash causes depression extending over many years.[8] The issue is delicate because the view that depression is an illness or disease of the medical sort does not sit in total comfort with its being caused by an event that happens to someone else. Lehman and colleagues focused on automobile accidents in Wayne County, Michigan, in which either a child under age 18, living at home, was killed, or in which a spouse aged 21 to 65 was killed. They used standardized criteria to exclude crashes in which the driver of the vehicle was responsible for causing the crash. They matched bereaved individuals to married controls from Wayne County who came to renew a driver's license, matching for gender, age, family income, education, and number and ages of children, producing 39 matched pairs of a bereaved individual and a control. Interviews were conducted four to seven years after the accident and included a variety of psychological measures of depression and well-being. The bereaved-minus-control differences in

measures of depression were substantial.[9] Reacting to the work of Bowlby and Freud, they concluded:

> The results suggest that sudden, unexpected loss of a spouse or child is associated with long-term distress . . . The results presented here suggest that the current theoretical approaches to bereavement may need to be reexamined . . . [because under these approaches] individuals are expected to return to normal role functioning and are not expected to experience distress several years after the loss has occurred.[10]

It is possible that the bereaved spouses and parents were different from the controls, because accidents are not entirely random events. Could such possible differences between the groups explain the difference in depressive symptoms many years later? Or is it more plausible that the sudden and unexpected loss of a spouse or a child does have long-term effects? Chapter 9, concerned with sensitivity analysis, will provide a quantitative dimension to these questions, a dimension informed by the empirical results, but it will not put the questions entirely to rest. What do you think? As I said at the outset, I found this study to be reasonably compelling.

Aliasing of Treatment with Something Thought to Be Innocuous or Controllable

The several examples just considered are built from a natural process that approximates random assignment to some degree. A less compelling but far more common type of natural experiment compares two groups of people who received different treatments, where there is no obvious reason to expect the two groups to be very different, but they might be different. The people on the left received the treatment, and those on the right received the control, but there was nothing special about the people on the left or the people on the right; in particular, no one moved to the left in an effort to receive the treatment. Common examples are similar geographic regions—say, U.S. states, Canadian provinces, or European Union member states—that implemented different treatments, or else people who were treated shortly before or shortly after a change in law or policy that altered the treatment that everyone receives.

The absence of an obvious reason to think that two groups are different falls well short of a compelling reason to think they are the same. Suppose that Sweden adopts a new policy that Norway declines to adopt. Can we estimate the effects of the policy by comparing outcomes in Sweden and Norway? After all, Sweden and Norway seem very similar to people who know little about them.[11] Upon inspection, we often find that people who live in adjacent U.S. states differ, on average, in age, education, income, and much more, even though we might have been ignorant of these differences before seeing them in data. Perhaps some or most of these visible differences can be removed, perhaps by stratification or matching as in Chapter 5, but it is always possible that some important difference was not measured and cannot be controlled in this way. It is common to try to shore up the claim that two groups are similar in relevant ways by showing that they are similar after treatment in terms of outcomes that the treatment should not affect, and differ only in terms of outcomes that the treatment should affect.[12]

In the technical literature on experimental design, two effects are said to aliased if the study design provides no way to tell them apart. If it is not possible to avoid aliased effects, then one might tolerate a design in which a plausible, potentially large and important effect is aliased with some other effect that seems implausible or small. If we compare two geographic regions that have adopted different policies whose effects we wish to estimate, then the effects of those policies are aliased with the other differences between those two regions. If we knew or thought or hoped that region was innocuous as a source of bias, then we might prefer to know that people received their treatments because of where they lived, as opposed to not knowing why people received their treatments. If we must tolerate nonrandom assignment of treatments, perhaps we would prefer what we hope is an innocuous source of bias to an unknown source of bias. "Nota res mala optima," wrote Erasmus, "an evil thing known is best."[13]

The Dutch famine of 1944–1945 was used to build a remarkable natural experiment with these issues in mind.

The Dutch Famine

An interesting and reasonably compelling natural experiment was conducted by Zena Stein, Mervyn Susser, Gerhart Saenger, and Francis Marolla in their

effort to estimate the effects of very poor prenatal nutrition on cognitive performance in adulthood.[14] During World War II, in September 1944, the Allies attempted to advance rapidly though Holland to reach Germany, aided by paratroops dropped on the town of Arnhem. The early part of this story became the plot for a popular movie, *A Bridge Too Far.* The natural experiment was built from the less familiar later part of this event.

The attempt to reach Germany failed. In reprisal, the Nazis imposed a transportation embargo on western Holland, causing a famine during the winter of 1944–1945. In the affected western areas, food rations were as low as 450 calories per day. Other areas of Holland were not affected. So a very rare thing happened. In an arbitrary part of a modern state, typical citizens experienced a famine, whereas in other parts of the same state they experienced no famine.

By knowing the date and place of birth of a child, one could know with reasonable accuracy whether the child's mother was affected by the famine while pregnant, and during which trimester of pregnancy she was affected. Additionally, there were children born earlier, too early to have prenatal exposure to the famine, and others born later, too late to have prenatal exposure. So the treatment—prenatal exposure to famine—is aliased with certain regions of Holland and certain periods of time; however, this type of aliasing seems relatively innocuous compared with the typical causes of extreme prenatal malnutrition. When male children reached the age of 18, they were required to take psychological and medical examinations as part of induction into the military. Stein and colleagues compared the results of these examinations for children with prenatal exposure to the famine to unaffected children from other parts of Holland or from time periods too early or late to have prenatal exposure to the famine. They found little or no difference in cognitive test performance at age 18 among males with and without prenatal exposures to the famine.

Can an Observational Study Be More Than "Reasonably Compelling"?

In discussing the study of depression by Lehman and colleagues and the study of the Dutch famine by Stein and colleagues, I spoke of them as "reasonably compelling." In describing the studies in this way, I mean that they

provided enduring evidence that must be considered and weighed, not casually disregarded, in any discussion of the topics that they investigated. To describe an observational study as reasonably compelling is to offer high and rare praise for the study, to distinguish it from run-of-the-mill studies that might reasonably be disregarded based on their poor design and inadequate analysis. Chapters 6 through 14 discuss aspects of what makes a few observational studies reasonably compelling. At present, we are only at the beginning, only considering the first aspect, namely, how closely an observational study resembles a randomized experiment. There are many other aspects.

The term "reasonably compelling" is intended to suggest, as one must, that even a reasonably compelling observational study may turn out, in light of subsequent research, to have reached an erroneous conclusion. Sometimes a reasonably compelling observational study prompts investigators to perform a randomized trial, and sometimes the trial does not support the conclusions of the observational study.[15] At other times, several reasonably compelling observational studies point in incompatible directions.[16] When ethical or practical constraints force scientists to rely on observational studies, it is not uncommon to see a decade or more of thrashing about, a decade or more of controversy, conflicting conclusions, and uncertainty. This can be true even when the studies themselves are well designed and executed.

Can an observational study be more than reasonably compelling? Arguably, it has happened once or twice, but reasonably compelling studies are rare to begin with.

In 1971, Arthur L. Herbst, Howard Ulfelder, and David C. Poskanzer published an observational study of diethylstilbestrol (DES) as a possible cause of vaginal cancer in young women, and arguably that study was more than reasonably compelling.[17] Diethylstilbestrol had been given to pregnant women in the hope that it would prevent complications of pregnancy. Today, it is believed to have caused vaginal cancer in the daughters of some of these women. Under normal circumstances, vaginal cancer is a very rare disease in young women. Herbst and colleagues studied eight cases of vaginal cancer in young women, finding that seven of the eight cases had in utero exposures to DES. Each case was matched to four other young women without vaginal cancer, born within five days at the same hospital, on the same type of service, ward or private. That is, there were 32 non-cases in total, four for

each of eight cases. None of the 32 non-cases had in utero exposures to DES. What makes the study compelling is the 7/8 versus 0/32 split; that is, an exceedingly rare disease is so strongly associated with an unambiguous treatment. In terms that will be made precise in Chapter 9, despite the small sample size, this study is extremely insensitive to unmeasured biases: only an enormous departure from a randomized experiment could produce this association if DES was not the cause.[18] There is nothing remarkable about the design or execution of the study by Herbst and colleagues. Rather, what is remarkable is the nature of the treatment, DES, the exceedingly rare disease of vaginal cancer, and their enormously strong association. It is the nature of this outcome and treatment that made this study more than reasonably compelling.

Though reasonably compelling observational studies are rare, it is within the power of investigators to design and execute reasonably compelling studies. If you want an observational study that is more than reasonably compelling, then you are going to need a little help from nature.

Aspects of Natural Experiments

Aspects of a Randomized Experiment That Might Be Present in a Natural Experiment

In a natural experiment, some aspects of a randomized experiment occur on their own. The relevant aspects vary from one study to the next. Most conspicuously, experiments are randomized, but some natural experiments are close to being randomized. Other aspects are presupposed in any actual experiment but may be absent in observational studies.

This section raises some of the ways a natural experiment may resemble, or fail to resemble, a randomized experiment.

Is Treatment Assignment Ignorable?

To what extent does treatment assignment in a natural experiment resemble randomization? As seen in Chapter 5, certain departures from complete randomization are corrected or repaired by stratification, matching, and suit-

able analysis. So we are especially interested in those departures from randomization for which repairs are not possible.

In Chapter 5, we saw that if treatment assignment were ignorable, then all the biases in estimating treatment effects come from measured covariates x_i, and by stratifying or matching for x_i these biases could be removed by appropriate analyses. In a slogan, ignorable treatment assignment means that people who look the same are the same. Ignorable treatment assignment means that two people, say, person i and j who look the same in terms of observed covariates, $x_i = x_j$, have the same probability of treatment, $0 < \pi_i = \pi_j < 1$. Stated a little more precisely, the probability, $\pi_i = \Pr(Z_i = 1)$, that person i receives treatment, $Z_i = 1$, may vary with the observed covariates, x_i, but among people with the same value of x_i, the probability π_i of treatment does not vary with their potential responses, (r_{Ti}, r_{Ci}).[19]

Is treatment assignment ignorable in a particular natural experiment? The answer will often be "yes" for natural experiments built from lotteries.[20] Waiting lists are more complicated. It may be hard to see how a waiting list can allocate treatments unfairly, but not everything that is hard to see is fair. People who live in different states or provinces are often seen to differ in terms of measured covariates, x_i, so how can we assert that they are the same in terms of unmeasured covariates u_i?

There is a quantitative dimension to departures from randomization or from ignorable treatment assignment. The departure may be small or large. Departing from ignorable assignment is descending a gradual slope, not falling off a cliff. We may leave the summit but remain close to it. The conclusions may unaffected by a small departure from ignorable assignment—they may be insensitive to such a departure—or they may be substantially altered by a small departure—they may be sensitive. Assessing this quantitative dimension is discussed in Chapter 9.

In building an observational study from a natural experiment, the investigator hopes to create a situation in which treatment assignment is either ignorable or nearly so.

Conditions Presupposed in a Randomized Experiment

Before one can conduct a randomized experiment, certain conditions need to be in place, and other conditions are almost invariably imposed by the

investigator. A natural experiment may achieve these same conditions even though randomization itself is not possible. The study by Lehman and colleagues of bereavement after traffic fatalities provides several illustrations.

In a randomized experiment, every person could receive either treatment or control; the turn of a coin decides. So it makes sense to speak of a causal effect, $\delta_i = r_{Ti} - r_{Ci}$, that compares the potential response of person i if assigned to treatment, r_{Ti}, or if assigned to control, r_{Ci}. Although fatal traffic accidents do not occur at random, they can happen to anyone who drives or rides in a car, so it is reasonable to speak of what would happen to me or you in this situation. On the other hand, it may be perfectly true and important that if Harry had been a woman, then his promotion at XYZ Corporation would have been blocked by the glass ceiling, yet it may also be perfectly unclear what we have in mind in changing Harry into a woman.[21] In a precarious situation, each person might actually receive either treatment, and the causal effect $\delta_i = r_{Ti} - r_{Ci}$ is well defined. The farther we move from a precarious situation, the less helpful it becomes to think of causal effects as comparisons of potential outcomes under alternative treatments. In seeking a natural experiment, the investigator seeks, among other things, a precarious situation.

In a randomized experiment, the investigator knows when the treatment started because the investigator started it. So the investigator knows which measures are pretreatment covariates and which other measures are posttreatment outcomes. The situation is similar in the nonrandomized bereavement study, because the fatal accident occurred abruptly at a well-documented time. In contrast, if one studied depression in response to the death of a spouse from Alzheimer's disease, the period of loss and bereavement may not have a well-defined beginning; it may begin before the spouse's death. In this sense, a sudden fatal accident more closely resembles an experimental treatment than does a death from Alzheimer's disease.

True experiments typically compare treatments that are very different.[22] In parallel, in the bereavement study, the sudden death of a spouse or a child living at home is very different from its absence. In this sense, the bereavement study resembles an experiment. One could increase the sample size, say, increase the number of bereaved people, but ruin the natural experiment by including people who lost elderly, distant relatives.

Concerning these presupposed elements in an experiment, Mervyn Susser wrote: "*Natural experiment* is a term I reserve for the observation of the effects of nonroutine, well-defined changes in environment. Such changes will best be events that are major, sharp, and out of the ordinary."[23]

The 2010 Chilean Earthquake and Post-traumatic Stress

Background: A Survey, an Earthquake, and Another Survey

This section introduces an example to which I will refer in later chapters. It concerns a study by José R. Zubizarreta, Magdalena Cerdá, and me about the possible effects on post-traumatic stress of the severe earthquake of magnitude 8.8 that struck Chile in 2010.[24] There have been many studies of post-traumatic stress after disasters.[25] An interesting feature of the literature about post-traumatic stress is the frequent observation that people have very different responses to extremely stressful situations such as major disasters, perhaps because the specific stresses people experience are objectively different, or perhaps because the people themselves differ in their resilience in stressful situations.[26] This heterogeneity of response, and the fact that the literature anticipates it, have methodological consequences.

Many studies of disasters collect data after the fact. Because the timing of disasters is not anticipated, investigators start asking questions after the disaster occurs. The worry about collecting data after the fact is that a person suffering from post-traumatic stress may remember the past differently than a person experiencing little or no stress. The worry is that the actual past may differ from the remembered past, and perhaps the errors of recall are different for people who were affected by the disaster. Expressing the same thought in the language of Chapter 1, in an after-the-fact study of the effects of a disaster, a covariate x_i is not truly a covariate—not truly a pretreatment measure—if we rely on recall of its value by someone whose thoughts have been affected by the disaster. The treatment we are studying, the disaster, may have affected not the past but how the past is recalled in the present. If that happened, then the recalled "covariate" for person i is actually an outcome, a quantity with two versions, say, x_{Ti} and x_{Ci}, that may be different because person i would recall the past differently if exposed to the disaster. When you mistake an outcome for a covariate, many things go wrong.[27]

Happily, the study by Zubizarreta and colleagues did not rely much on recall. Shortly before the earthquake, Chile's Ministry of Planning and Cooperation had completed its national socioeconomic survey, called the CASEN. This ministry had the quickness of mind and wisdom to draw a subsample of the respondents to the CASEN and to interview them again

after the earthquake, creating rare longitudinal data before and after a major disaster. Longitudinal data refer to data collected on the same people at two or more times; with few exceptions, longitudinal data are preferable to relying upon recall of the past.

The second interview included the Davidson Trauma Scale, a rating scale for post-traumatic stress.[28] In the discussion here, the Davidson Trauma Scale is the primary outcome. The scale was built to resemble the discussion of post-traumatic stress in the American Psychiatric Association's *Diagnostic and Statistical Manual of Mental Disorders IV.* Specifically, two questions were asked about each of 17 symptoms. One question asked about the frequency of the symptom, scoring frequency from 1 = "not at all" to 5 = "every day." The second question asked about the severity of the symptom, scoring from 1 = "not at all distressing" to 5 = "extremely distressing." For example, the question "Do you have difficulty falling asleep and remaining asleep?" was scored 1 to 5 for frequency and 1 to 5 for severity. In the end, each person is scored twice on 17 questions, or $34 = 2 \times 17$ times, and can have a score as low as $34 = 1 \times 2 \times 17$ by answering 1 every time, or as high as $170 = 5 \times 2 \times 17$ by answering 5 every time. A score of 34 indicates no reported symptoms of post-traumatic stress, and a score of 170 indicates the highest possible score that this scale can record. To what extent, if any, does exposure to an earthquake increase reports of post-traumatic stress symptoms?

The Davidson Trauma Scale describes one real thing, namely, what people say in describing their symptoms in response to a highly structured set of questions. That one real thing might differ from other real things, such as ratings that a psychiatrist might provide, because the psychiatrist might respond to aspects of the person not captured by structured questions. Moreover, different psychiatrists might rate the same person differently. Similarly, a spouse might describe a person differently than the person describes himself or herself; here, the spouse's description is a third real thing, perhaps different from both the Davidson Scale and the psychiatrist's rating. Because only the Davidson Scale is available here, this study is asking whether what people say in such a structured interview was affected by the earthquake.

A Modern Matched Comparison

Zubizarreta and colleagues formed 2,520 matched pairs of two people, one living in a region of Chile severely shaken by the earthquake and the other

living in a region far removed from and little shaken by the earthquake. The pairs were matched for 46 measured covariates x_i describing people before the earthquake, including demographic characteristics such as age, gender, marital status, and ethnic group, socioeconomic measures, and indicators of the nature and quality of housing, health, disability, employment, education, and health insurance. Perhaps unsurprisingly, people who live in different regions of Chile differ in terms of many of these covariates, including employment, income, education, ethnic group, and health insurance.

As seen in Table 5.6, there are not enough people on earth to match exactly for 46 covariates. Nonetheless, matching produced a treated group and a control group that were similar in terms of these measured covariates. To reiterate a point from Chapter 5, finding pairs of people who look the same on 46 covariates is very difficult, but producing two groups of people with similar distributions of 46 covariates is often a practical problem that technical tools can solve. For example, after matching, the exposed and control groups each had (i) 406 people aged 55 to 64 years, and similarly for other age categories; (ii) 210 people from the indigenous ethnic group; (iii) 122 people whose self-rated health was "poor"; and (iv) 1,738 people whose housing quality was rated as "acceptable."

Table 6.1 shows the distributions of age and years of education in the matched exposed and control groups. In Table 6.1, the median cuts the distribution in half, so half the exposed people are 46 years old or younger, and half the controls are 47 years old or younger. The quartiles cut the distribution in quarters; so in both exposed and control groups a quarter of the people are 35 years old or less, and so on. After matching, the distributions of age

Table 6.1. Distribution of age and years of education in 2,520 matched pairs of one individual exposed to the Chilean earthquake and a control remote from exposure

	Minimum	Lower quartile	Median	Upper quartile	Maximum
			Age in years		
Exposed	15	35	46	61	93
Control	15	35	47	60	97
			Years of education		
Exposed	0	6	10	12	20
Control	0	6	10	12	20

and years of education are quite similar. In the matched comparison, these covariates were balanced, as were all 46 covariates. In the original article and in any competent report of a matched comparison, there is extensive information similar to Table 6.1 showing that the match succeeded in balancing the observed covariates.

The match was constructed using several modern techniques that will be discussed in Chapter 11. These techniques included an estimated propensity score, as discussed in Chapter 5, various "fine balance" constraints, and a within-pair distance that was minimized by integer programming.[29] The 46 covariates were used to estimate the propensity score—that is, the probability that each person i lives in a severely shaken region of Chile based on the observed covariates x_i—and then people with similar propensity scores were paired. Fine balance means forcing perfect balance for certain variables without constraining who is matched to whom. For example, housing quality was finely balanced so that exactly 1,738 people had "acceptable" housing in both the exposed and control groups. Additionally, an attempt was made to pair people who were as similar as possible in terms of x_i, recognizing from Table 5.6 that there are severe limits to how close such a pairing can be.

Post-traumatic Stress Outcomes

Table 6.2 shows the distribution of Davidson Trauma Scores in the exposed and control groups. A score of 34 is the lowest possible score, and the median or middle score of the 5,040 scores was 37. The upper quartile of the 5,040 scores was 54, so 75% of the scores were 54 or less. The 90th percentile of the 5,040 scores was 79, so 90% of the scores were 79 or less. The maximum possible score was 170. Notably, in Table 6.2 many people in the exposed group showed little or no sign of post-traumatic stress—710 had scores of 37 or less—but scores of 80 and above were much more common in the exposed group than in the control group, 454 versus 35.

Table 6.3 displays the same data with greater care. Table 6.3 resembles Table 5.5 in that both tables count matched pairs, not people, so the total count in Table 6.3 is 2,520 pairs, not the 5,040 people in Table 6.2. The row of Table 6.3 indicates the trauma score for the exposed individual in a pair, and the column indicates the trauma score for the control in the same

Table 6.2. Distribution of Davidson trauma scores in 2,520 matched pairs of one
individual exposed to the Chilean earthquake and a control remote
from exposure

Trauma score	34–37	38–54	55–79	80–170	Total
Exposed	710	772	584	454	2,520
Control	1,822	521	142	35	2,520
Total	2,532	1,293	726	489	5,040

The minimum possible score is 34, the median is 37, the upper quartile or 75th percentile is 54, the 90th percentile is 79, and the maximum possible score is 170. The table counts 5,040 individuals.

Table 6.3. Joint distribution of Davidson trauma scores for 2,520 matched pairs of
one individual exposed to the Chilean earthquake and a control remote
from exposure

		Trauma score for control				
		34–37	38–54	55–79	80–170	Total
	34–37	532	132	37	9	710
Trauma score	38–54	569	149	43	11	772
for exposed						
	55–79	417	128	30	9	584
	80–170	304	112	32	6	454
	Total	1,822	521	142	35	2,520

The table counts 2,520 pairs, not 5,040 individuals.

pair. The row and column totals at the far right and bottom of Table 6.3 equal the rows of Table 6.2. Had Table 6.3 come from a randomized experiment in which one person in each pair was picked at random for exposure, the pattern in Table 6.3 would be unambiguous: it would say that the earthquake caused substantial symptoms of post-traumatic stress, but many exposed people experienced few or no symptoms. In particular, there are extensions of McNemar's test, used in Table 5.5, for tables such as Table 6.3 with more than two outcome categories. Like McNemar's test, these extensions look at discordant pairs, although there are now many types of discordant pairs.

Do Some Matched Pairs Provide More Information
Than Others about Causal Effects?

Think about it this way: concordant pairs, counted on the diagonal, are analogous to a tie or draw in a treatment–control contest, but a discordant pair has a winner and a loser. Table 6.3 records the results of 2,520 contests between treatment and control. In Table 5.5, you win or you lose or you tie, as in chess. In Table 6.3, wins and losses come with points or scores, as in football, and we can ask, By how much did you win? As in sports, a win is a win, but a big win convinces us that the winning team is stronger. The information in a big win is different from the information in a close win.

In Table 6.3, we care about wins and losses, but also about the magnitudes of the victories. We compare the frequency of 1-to-2 losses to the frequency of 2-to-1 victories, but we distinguish these from 1-to-4 losses and 4-to-1 victories. In a randomized experiment with no treatment effect, the frequency of 1-to-2 losses would be similar to the frequency of 2-to-1 victories, and the frequency of 1-to-4 losses would be similar to the frequency of 4-to-1 victories; the differences would be due to chance, not to the strength of the opponents. We recognize a treatment effect by a preponderance of wins, but we judge big wins to be more convincing.

For example, if we ignore all of Table 6.3 except the four cells in the first two rows and columns, then the resulting 2×2 table resembles Table 5.5 in form, and it contains $701 = 569 + 132$ pairs discordant for no symptoms versus slight symptoms, for scores in [34, 37] versus [38, 54]. If there were no effect of the earthquake in a randomized experiment, we expect these 701 discordant pairs to behave like 701 flips of a fair coin, but in fact $569 / 701 = 81\%$, not 50%, showing higher symptoms for the exposed person in the pair. Of course, Table 6.3 is not from a randomized experiment, so we cannot rely on reasoning of this kind.

Look at two corner cells of Table 6.3, where there are $313 = 304 + 9$ discordant pairs in which one person was fine and the other had severe symptoms; that is, one person has a trauma score of 80 or more, and one person has a score of 37 or less. In $304 / 313 = 97\%$ of these pairs, it was the exposed individual, not the control, who had severe symptoms, the trauma score of 80 or more. Many exposed individuals showed little or no sign of posttraumatic stress, and many people in both groups reported a few mild symptoms now and then. In the 313 extreme pairs with one person who is fine

and another who is experiencing substantial symptoms, the symptoms track the exposure quite closely, 97% of the time.

A statistical analysis that pretends Table 6.3 is from a randomized experiment pays close attention to all of the discordant counts, both the 569 / 701 = 81% split for [34, 37] versus [38, 54] and the 304 / 313 = 97% split for [34, 37] versus [80, 170]. An analysis that acknowledges that people are not randomly assigned to live in different parts of Chile needs to ask whether the pattern in Table 6.3 could be due to some bias, some failure to match for a covariate that was not measured in the survey. Actually, it would take a spectacularly large bias from an unmeasured covariate to produce Table 6.3 if the earthquake had no effect; however, the 304 / 313 = 97% split for [34, 37] versus [80, 170] is much more important to this claim than is the 569 / 701 = 81% split for [34, 37] versus [38, 54]. This issue will be examined in detail in Chapters 9 and 10.[30]

Taking Stock

A natural experiment is an attempt to find in the world a situation having some of the desirable attributes of a randomized experiment. Natural experiments vary considerably in the degree to which they do approximate randomized experiments. One thinks first about random or ignorable treatment assignment. Less of an achievement, but still important, are settings in which a routine source of bias is absent, even if we cannot be sure that all sources of bias are absent. Also important are well-defined treatments that are substantially different, beginning abruptly at a well-documented moment.

Elaborate Theories

[We should] trust rather to the multitude and variety of . . . arguments than to the conclusiveness of any one. [Our] reasoning should not form a chain which is no stronger than its weakest link, but a cable whose fibers may be ever so slender, provided they are sufficiently numerous and intimately connected.

—CHARLES SANDERS PEIRCE[1]

All human errors are impatience, the premature breaking off of what is methodical.

—FRANZ KAFKA[2]

What Are "Elaborate Theories"?

Cochran's Discussion of Fisher's Advice

In an observational study, in a study of treatment effects without random assignment of treatments, an association between treatment received and observed outcome is ambiguous: it could reflect an effect caused by the treatment, or an unmeasured bias in the way treatments were assigned, or a combination of the two. In what ways can this ambiguity be reduced?

Sir Ronald Fisher invented randomized experimentation, and William G. Cochran first presented observational studies as a topic in statistics defined by its contrast with randomized experiments. How did they perceive the problem of recognizing effects actually caused by treatments when treatments are not randomly assigned? Cochran wrote,

First, as regards planning. About 20 years ago, when asked in a meeting what can be done in observational studies to clarify the step from association to causation, Sir Ronald Fisher replied: "Make your theories elaborate." The reply puzzled me at first, since by Occam's razor the advice usually given is to make theories as simple as is consistent with the known data. What Sir Ronald meant, as the subsequent discussion showed, was that when constructing a causal hypothesis one should envisage as many different consequences of its truth as possible, and plan observational studies to discover whether each of these consequences is found to hold . . . This multi-phasic attack is one of the most potent weapons in observational studies. In particular, the task of deciding between alternative hypotheses is made easier, since they may agree in predicting some consequences but will differ in others . . . The combined evidence on a question that has to be decided mainly from observational studies will usually consist of a heterogeneous collection of results of varying quality, each bearing on some consequence of the causal hypothesis . . . [The investigator] cannot avoid an attempt to weigh the evidence for and against, since some results are so vulnerable to bias that they should be given low weight even if supported by routine tests of significance.[3]

Suppose that a causal theory makes several predictions—that it is elaborate. Then there are several associations predicted by the theory, and each can be checked against data. Suppose that a skeptic challenges one of these associations, the skeptic's claim being that the association is produced by a particular bias in who is treated, not by any effect caused by the treatment. If a second association predicted by the causal theory cannot be explained by the skeptic's first postulated bias, the skeptic must justify continued skepticism by postulating a second bias to explain the second association, and so on. The weighing of evidence for or against a causal theory may become clearer as there is more evidence to weigh.

Let us consider a quick, simple example of the use of an elaborate theory in an observational study, before returning to the general topic.

A Simple Example: Does a Parent's Occupation Put Children at Risk?

Adults are often exposed to occupational hazards at work, and a variety of laws regulate these exposures to ensure the safety of workers. Might a parent

be exposed to an occupational hazard in such a way that his or her child is also exposed, even if the child never enters the workplace?

David Morton and colleagues asked whether parents who work in an industry using lead might bring lead home in their clothes and hair, thereby exposing their children.[4] They measured the level of lead in the blood of 34 children with a parent who worked in a battery manufacturing plant in Oklahoma that used lead in the production of batteries. As it turned out, all these parents were fathers. Morton and colleagues found control children whose parents did not work in the battery plant. They paired each treated child to a control child from a different household whose age differed by at most one year and who lived close by in the same neighborhood. If a treated child lived in a home facing a major road, the control child was selected from the same side of the same road. If the treated child lived in an apartment complex, the control child came from the same complex. This matching was intended to ensure that treated and control children faced similar levels of environmental lead at home, say, from automobile exhaust or nearby industrial pollution. For both groups, they measured levels of lead in the children's blood, recorded in micrograms of lead per deciliter (μg/dl) of blood. As I write in 2016, the U.S. Centers for Disease Control and Prevention says, "Experts now use a reference level of 5 micrograms per deciliter to identify children with blood lead levels that are much higher than most children's levels."[5] At the time of Morton and colleagues' study, in 1982, a higher level was often used, 30 μg/dl.

Additional information was collected. First, some workers held jobs in the battery plant constantly exposing them to lead, whereas others had limited exposure. Using information from the plant manager, the father's exposure to lead was graded high, medium, or low depending upon the specific job the father performed. There were 19 fathers with high exposure, 7 with medium exposure, and 8 with low exposure. Second, an interviewer questioned homemakers about occupational hygiene, and on that basis the workers were graded as having good, moderately good, or poor hygiene. For instance, a lead worker had good hygiene if he showered, shampooed, and changed shoes and clothes at work before going home, whereas changing clothes without showering was considered moderately good. As one might expect, fathers with little exposure to lead at work rarely showered and changed before going home, so we will look at the hygiene for 19 fathers with high exposure, where 13 had poor hygiene, 3 had moderately good hy-

giene, and 3 had good hygiene. For plotting purposes, the $6 = 3 + 3$ fathers
with moderately good or good hygiene are combined into a single group
called OK hygiene.

What is the elaborate theory? If a father's exposure to lead has an effect
on the lead level of his children, then we expect (i) higher levels of lead in
the blood of treated children than in matched control children, (ii) higher
levels of lead in the blood of children whose fathers have higher exposure to
lead at the battery plant, and (iii) higher blood levels if a high-exposure father
practices poor hygiene. Additionally, (iv) the lead exposure of the father of
a treated child should not predict the blood lead level of the control child to
which the treated child is paired. One way prediction (iv) could fail is if high-
exposure fathers live in a poor neighborhood near the battery plant, and
people who live near the battery plant are exposed to air pollution from the
plant, and matched control children from the same neighborhood are also
exposed to the same air pollution.

The results will be displayed using boxplots, a widely used graphical dis-
play invented by John W. Tukey.[6] Figure 7.1 illustrates a boxplot with fic-
tional data about a variable Y. A boxplot has a central box, vertical lines that
continue up and down from the box, and may contain individual points.
The central horizontal line in a boxplot is placed at the median, the middle
value in sorted data, so half the values of Y are above this median line and
half are below. The upper horizontal line in the box is at the upper quartile,
so one quarter of the values of Y are above the upper horizontal line, and
three quarters are below. In parallel, the lower horizontal line in the box is
at the lower quartile, so one quarter of the values of Y are below the lower
horizontal line, and three quarters are above. Saying the same thing differ-
ently: a quarter of the Y's are below the box, a quarter are in the lower part
of the box, a quarter are in the upper part of the box, and a quarter are above
the box. Points judged extreme by a certain conventional standard are la-
beled individually, and there are five such points in Figure 7.1: four at the
top, and one at the bottom. The vertical lines extend upward and downward
from the box to the last value of Y that is not judged extreme.

As Tukey emphasized, boxplots convey quite a bit of information about
a variable Y. The median line shows the typical value of Y. The box outlined
by the upper and lower quartiles indicates a range that includes a central
half of the Y's, thereby indicating how much the Y's typically vary. The rest
of the plot tells us about atypical values of Y.

Fictional Example of Tukey's Boxplot
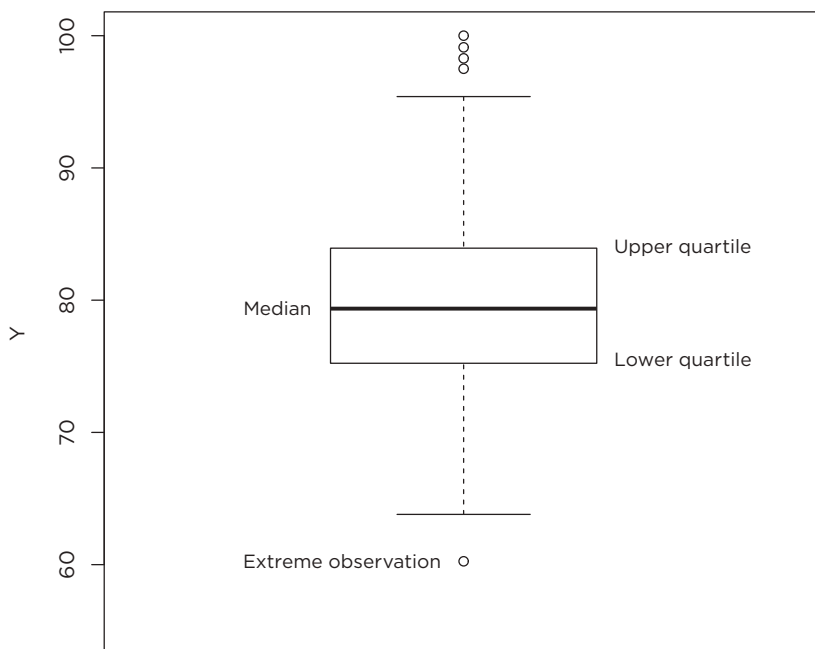
Figure 7.1. A boxplot for a fictional variable Y. Half of the values of Y are above the median (the middle line in the box), and half are below. A quarter of the values of Y are above the upper quartile (the upper line in the box), and three quarters are below. A quarter of the values of Y are below the lower quartile (the lower line in the box), and three quarters are above. Individual values of Y judged to be extreme by a certain standard are plotted as individual points.

Figure 7.2 checks the elaborate theory against the data. In Figure 7.2, a treated child is one whose father works in the battery plant, whereas a matched control child is of similar age and lives in the same neighborhood. Panel (a) on the upper left of Figure 7.2 compares the lead levels in the blood of treated children and matched control children. The blood lead levels are much lower for control children. Panel (c) on the lower left of Figure 7.2 focuses on the treated children, distinguishing among them on the basis of their fathers' levels of exposure to lead at the battery plant. If a father has high exposure, his child is more likely to have a higher level of lead in the blood. The level of exposure to lead of the treated child does not predict the
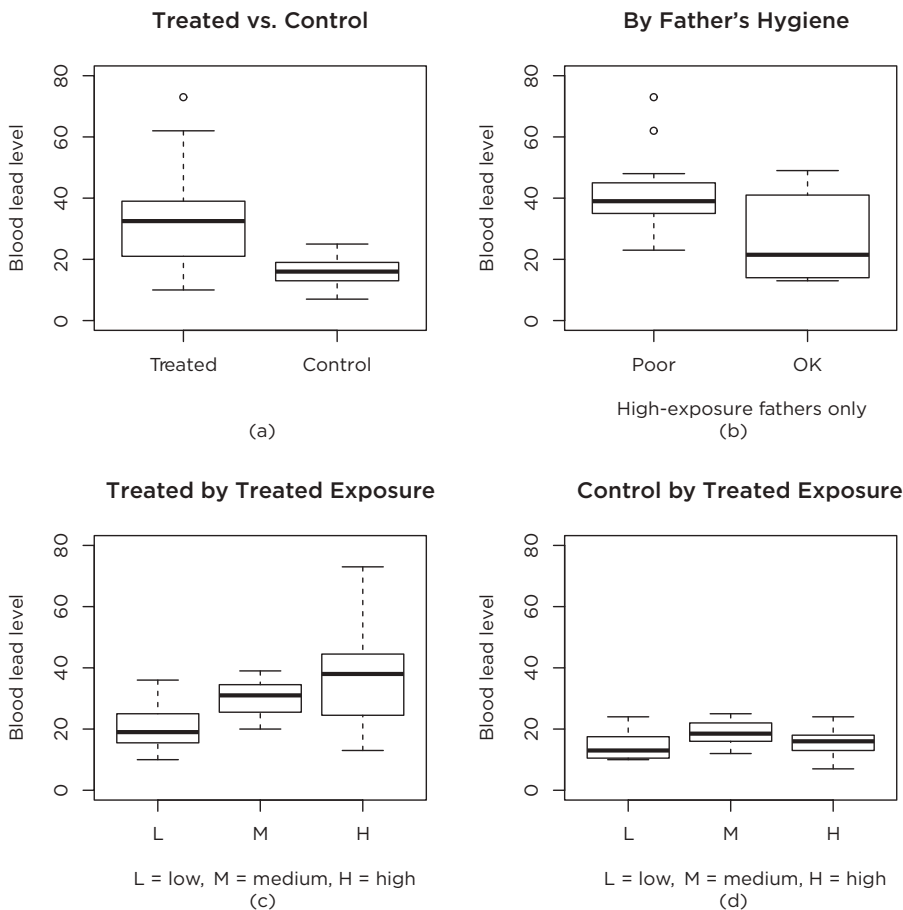
Figure 7.2. Checking an elaborate theory. Lead levels in the blood of children, μg/dl. Panel (a) compares children whose fathers worked in the battery plant to matched control children of similar age from the same neighborhood. Panel (c) separates treated children based on their fathers' levels of exposure to lead at the battery plant. Panel (d) separates control children based on the level of exposure of the father of their pair-matched treated child. Panel (b) looks only at children of fathers with high exposure to lead, separating them on the basis of the father's hygiene before leaving the factory, either poor or OK, where the OK group merges three fathers with good hygiene and three fathers with moderately good hygiene.

level of lead in the blood of the matched control child in panel (d), so it would be difficult to attribute the pattern in panel (c) to differences among neighborhoods, because the neighborhoods are the same in panels (c) and (d). Finally, in panel (b), if a father has high exposure at the battery plant but practices better hygiene when leaving the plant, then the child's lead level is lower. In brief, each prediction of the elaborate theory agrees with the observed data.

Figure 7.2 does not ensure that the higher blood lead levels for treated children were caused by their fathers' exposure to lead at work. However, it is not easy to think of something else that could produce all of the patterns in Figure 7.2: lower lead levels for controls, lower lead levels with low exposure, and lower lead levels with better hygiene.

Figure 7.2 is a particularly simple example of an elaborate theory checked against data within a single study. Let us return to the general discussion of elaborate theories.

Aspects of Elaborate Theories

The Logic of Elaborate Theories

An elaborate theory makes extensive predictions about what will be observed, so it is less likely to be true than a theory that makes fewer predictions, and it is more likely to be contradicted by observed data. Are these desirable features of a theory? If so, why are they desirable?

Philosophers of science have often argued that they are desirable features of a scientific theory. In an essay, "On Selecting Hypotheses," Charles Sanders Peirce wrote, "But if I had the choice between two hypotheses . . . I should prefer . . . [the one which] would predict more, and could be put more thoroughly to the test . . . It is a very grave mistake to attach much importance to the antecedent likelihood of hypotheses . . . Every hypothesis should be put to the test by forcing it to make verifiable predictions."[7] Similarly, Karl Popper wrote,

Theories may be more, or less, severely testable; that is to say, more, or less, easily falsifiable. The degree of their testability is of significance for the selection of theories . . . We should try to assess what tests, what trials, [the

theory] has withstood . . . It is not the number of corroborating instances which determines the degree of corroboration as the severity of the various tests to which the hypothesis can be, and has been, subjected. But the severity of the tests, in its turn, depends upon the degree of testability . . . We try to select for our tests those crucial cases in which we should expect the theory to fail if it is not true.[8]

Mechanisms by Which the Effect Is Produced

Elsewhere, Cochran wrote, "A claim of proof of cause and effect must carry with it an explanation of the mechanism by which the effect is produced. Except in cases where the mechanism is obvious and undisputed, this may require a completely different type of research from the observational study that is being summarized."[9]

The claim that a treatment produces its effects by the operation of a particular mechanism is an elaboration of the causal theory, one that creates additional predictions and hence additional opportunities to check the theory's predictions against observation. Similarly, an argument that a proponent of a policy has offered claiming that a treatment could or will have certain effects creates opportunities to empirically study the elements of that argument. In parallel, elaboration of a skeptical counterclaim saying that an observed association is produced by bias creates opportunities to check the counterclaim against observation.[10]

In studying whether smoking causes lung cancer, an elaborate theory may predict that (i) smokers will develop lung cancer more often than nonsmokers in observational studies of people,[11] (ii) laboratory animals experimentally exposed to tars in cigarettes will develop cancer,[12] and (iii) the autopsied lungs of smokers who died of something other than lung cancer will exhibit cellular damage similar to that of individuals who died of lung cancer and unlike the lungs of nonsmokers.[13] These very different types of research each have weaknesses—smoking is not randomized, and mice are not people—so each comparison may be unconvincing on its own, but agreement between studies with very different weakness may be compelling.

Quantitative observational studies of many people may complement narrative, ethnographic, or qualitative studies of a few people.[14] A covariate that is not measured in a large quantitative study may be available for the

asking in an ethnographic study. Here, again, the weaknesses of one type of investigation may differ from those of another. Agreement among studies with very different weaknesses gradually makes it more difficult to attribute a shared conclusion to bias produced by their weaknesses.

How to Elaborate a Causal Theory

> What goes on in science is not that we try to have theories that accommodate our experiences; it's closer that we try to have experiences that adjudicate among our theories.
>
> —JERRY FODOR[15]

In thinking about how best to elaborate a causal theory, what considerations are important? The quote from Cochran at the beginning of the chapter makes a key suggestion: with an elaborate theory, "the task of deciding between alternative hypotheses is made easier, since they may agree in predicting some consequences but will differ in others." Suppose that a causal theory has met with some challenge, or is evidently open to some challenge— that is, some reasonably specific counterclaim explaining how the observed association between treatment and outcome is produced by a biased comparison, not an effect caused by the treatment. We are especially interested in an elaboration of the causal theory such that the causal theory and this counterclaim make different predictions about something we can observe. In this way, the elaboration of the causal theory helps to adjudicate between the causal theory and a specific counterclaim.

Let us consider an example.

Effects on Crime Rates of Restricting Handgun Purchases

An interesting example is from a study by Garren Wintemute, Mona Wright, Christiana Drake, and James Beaumont of the possible effects that restrictions on handgun purchases might have on crime rates.[16] Table 7.1 is adapted from their table 2. U.S. federal law restricts the purchase of a handgun by people who have been convicted of a felony. In 1991, the state of California went further, prohibiting for 10 years the purchase of a handgun by a person

Table 7.1. Crime rates in California, 1989–1991, among people who tried to purchase a handgun following a conviction for a violent misdemeanor

		Events per 100 person-years of observation		
Handgun purchase	*Number of people*	*Any crime*	*Gun or violent crime*	*Nonviolent crime with no gun*
Denied	927	14.1	8.0	9.3
Approved	727	15.5	9.9	8.6

convicted of certain violent misdemeanors, including assault, resisting arrest, and brandishing a firearm. Wintemute and colleagues compared two groups of people who would have qualified in 1991 for this restriction on handgun purchases. One group had attempted to make a handgun purchase in 1991, and because of the new restriction the purchase was denied. The second group attempted to purchase a handgun in 1989 or 1990, and because the new restriction was not in effect, the purchase was approved. In other words, the two groups, denied and approved, were the same in the sense that the change in law changed their ability to legally purchase a handgun, but they differed in when they attempted to make the purchase. In a table not unlike Table 1.1 for the ProCESS Trial, Wintemute and colleagues compared these two groups in terms of measured covariates, arguing that the groups were similar in terms of gender, age, race, and ethnicity, number of prior convictions, and number of convictions for gun or violent crimes. They then compared outcomes—namely, arrests for crimes committed after the purchase attempt.

There is a technical issue in Table 7.1 that I will mention but not discuss in detail. People were observed for different periods of time. If Harry was observed for a longer period of time than Sally, then Harry is more likely than Sally to be observed to commit a crime, just because we kept an eye on him for a longer time. We do not want this trivial issue to be confused with any possible effects of the change in law. For this reason, Wintemute and colleagues used several technical tools to address unequal periods of observation, the most elementary of these being evident in Table 7.1. Specifically, arrest rates are not per person but per person-year of observation. If Harry is observed for two years, he contributes two person-years to the study. If Sally is observed for one year, she contributes one-person year to the study.

If Harry is arrested twice in his two years and Sally is arrested once in her one year, then the arrest rates for Harry and Sally are the same, one arrest per person-year. The rates in Table 7.1 are high: a rate of 15 arrests per 100 person-years for Harry means, roughly speaking, a 15% chance Harry is arrested each year. Remember that the denied and approved groups both consist of people who had been convicted previously of certain violent misdemeanors.

The comparison of denied and approved groups suffers from one evident defect: the denials all occurred in a later year than the approvals. If criminal activity shifted greatly over the period from 1989 to 1991, then that shift might be confused with an effect of the change in law. For instance, changes in the unemployment rate or the activities of the police from 1989 to 1991 might affect whether a person is inclined to commit a crime. If Harry and Sally are out of work and short of cash in a particular year because of the sour economic situation in that year, then they might be more inclined to commit a crime that yields cash. Perhaps the ups and downs in criminal activity reflect the changing economic situation, not the change in law restricting handguns. We want an elaboration of the causal theory so that a sour economy predicts one thing will happen, but an effect of handgun restrictions predicts something else will happen.

The causal theory that handgun restrictions affect crime rates permits an obvious elaboration: restrictions on handguns should reduce specifically crimes for which possession of a handgun is relevant. If Harry is inclined to demand someone's wallet at gunpoint, then not being in possession of a gun at that delicate moment might prove inconvenient. This is less of an inconvenience for Sally, who is inclined to pick someone's pocket. A rise in the unemployment rate in a particular year might increase the rate of crimes committed for monetary gain, but that tendency seems unlikely to be restricted to crimes for which a gun is relevant. In the elaborate causal theory, an effect of the change in law restricting access to guns predicts a different pattern of associations than does a sour economic situation.

Table 7.1 and other analyses by Wintemute and colleagues show a lower rate of arrests for gun and/or violent crime in the group whose gun purchase was denied, but not a lower rate of nonviolent crime without a gun. The visible pattern in Table 7.1 is easier to explain as an effect of restrictions on gun purchases, harder to explain in terms of a sour economy. The orig-

inal study should be consulted for further analyses that involve, as I have mentioned, some additional technical detail.

Could Table 7.1 still be produced by a biased comparison? Yes, but not by a bias that affects crimes of all kinds in a similar way. If the police had decided in 1991 to crack down on violent gun crime and ignore pickpockets, then that shift might depress violent gun crime in the denied group without depressing nonviolent crime in the denied group, consistent with Table 7.1, even if there were no effect of the restrictions on handgun purchases. Of course, it would be easy to check in other ways whether such a shift in police behavior took place.

An elaborate theory is most valuable if it helps to adjudicate between a treatment effect and some plausible counterclaim denying such an effect.

Elaborate Theories and Tests of Ignorable Treatment Assignment

Recall the situation discussed in Chapter 5. In Chapter 5, you have a treated and a control condition, $Z_i = 1$ or $Z_i = 0$, an observed outcome, R_i, that equals $R_i = r_{Ti}$ if $Z_i = 1$ or $R_i = r_{Ci}$ if $Z_i = 0$, and an observed covariate, x_i. In this situation, you would have what you need for causal inference if treatment assignment were ignorable, but you have no way to check whether indeed treatment assignment is ignorable. How can you check whether treatment assignment is ignorable?

An elaborate theory changes this situation. With such a theory, there are things you could observe that would force you to abandon either the elaborate theory or the claim that treatment assignment is ignorable. With firm commitment to an elaborate theory, you can test ignorable treatment assignment. Moreover, framed as a statistical test of hypotheses, it is possible to ask about the properties of the test, for instance, whether the test is likely to detect failures of ignorable assignment when particular types of bias are present.[17]

The Crossword Analogy

Cochran's quoted discussion of elaborate theories spoke of weighing evidence from "a heterogeneous collection of results of varying quality, each bearing

on some consequence of the causal hypothesis . . . some [of which] are so vulnerable to bias that they should be given low weight even if supported by routine tests of significance." Here, there are many strands of evidence of uneven quality, none strong enough to be compelling on its own, with some strands intersecting in ways that provide mutual support while others clash in ways that bar the emergence of a coherent picture. So a process is needed that appraises mutual support among studies in the presence of conflict among studies, a process that sets aside bits of evidence to see whether a strongly supported, coherent picture emerges from what remains.

In a related context, the philosopher Susan Haack suggested an analogy between the development of scientific knowledge and the solving of a crossword puzzle:

> The model is not . . . how one determines the soundness or otherwise of a mathematical proof; it is, rather, how one determines the reasonableness or otherwise of entries in a crossword puzzle. This model is more hospitable to a graduational account . . . The crossword model permits pervasive mutual support, rather than, like the model of a mathematical proof, encouraging an essentially one-directional conception . . . How reasonable one's confidence is that a certain entry in a crossword is correct depends on: how much support is given to this entry by the clue and any intersecting entries that have already been filled in; how reasonable, independently of the entry in question, one's confidence is that those other already filled-in entries are correct; and how many of the intersecting entries have been filled in.[18]

Haack is making two points. First, much of the conviction that a penciled-in crossword puzzle is correct stems from appropriate intersections of entries, rather than the strength of the individual clues that support individual entries. In parallel, in science, conviction often results from the appropriate intersection or agreement of several or many inconclusive studies with different vulnerabilities. The number of studies, their sample sizes, and their levels of statistical significance are not the critical issues. The critical issue is whether the vulnerability that makes one study doubtful is absent from another study. Observational studies of people, experimental studies of laboratory animals, and experimental studies of interactions among biomolecules are each vulnerable to error in determining the effects of treatments on people, but their vulnerabilities are very different.

Second and more subtly, Haack wishes to exhibit the possibility of mutual support without vicious circularity. Suppose I can deduce A from B, and I can also deduce B from A; then, to believe A and B are both true on that basis alone would be to err by vicious circularity, because logically A and B could both be false. Two entries in a crossword may meet appropriately yet both be incorrect. In the quote above, Haack asks, "How reasonable, independently of the entry in question, one's confidence is that those other already filled-in entries are correct"? Haack suggests that mutual support among entries in a crossword is possible without vicious circularity by a process of demarcation. Demarcation means setting aside a specific part of the evidence when appraising another part.[19] When using the evidence supporting A and its appropriate intersection with B in appraising the evidence available for B, it is appropriate to leave aside the evidence that B provides for A. In parallel, when using the evidence supporting B and its appropriate intersection with A in appraising the evidence available for A, it is appropriate to leave aside the evidence that A provides for B. In a crossword puzzle in which 1-down intersects 3-across, we may ask what evidence we have for our tentative solution to 1-down apart from its appropriate intersection with 3-across, and we may ask what evidence we have for our tentative solution to 3-across apart from its appropriate intersection with 1-down. We might conclude that the only compelling evidence for our tentative solution to 1-down comes from its appropriate intersection with 3-across, and the only compelling evidence for our tentative solution to 3-across comes from its appropriate intersection with 1-down; then, we would regard these two solutions as very tentative. Alternatively, we might conclude that our tentative solution to 1-down is supported by strong evidence apart from its appropriate intersection with 3-across, and that 3-across is supported by strong evidence apart from its appropriate intersection with 1-down; then, the appropriate intersection of these two entries provides additional support that they are each correct.

In appraising evidence from several observational studies, or from one study with several comparisons, we might ask: What evidence is available in support of a particular conclusion apart from those studies that suffer from a particular vulnerability? We might ask this question repeatedly, for different vulnerabilities.

The studies of fetal alcohol syndrome provide an example.

Fetal Alcohol Syndrome

> Make ready a feast of the princes. There it is your pleasure to eat the roast
> flesh, to drink as much as you please the cups of the wine that is sweet as
> honey.
>
> —HOMER, *The Iliad*[20]

The *Iliad* and intoxication are older than the written word, but the substantial effects of alcohol on the developing human fetus received systematic investigation in the latter half of the twentieth century.[21] At high, steady doses, the effects of alcohol on human fetal development are difficult to miss. The initial reports were case series of children born to severely alcoholic mothers with no comparison group.[22] As one case-series began, "Eight unrelated children of three different ethnic groups, all born to mothers who were chronic alcoholics, have a similar pattern of craniofacial, limb, and cardiovascular defects associated with prenatal-onset growth deficiency and developmental delay." The greatest harms were done to developing brains, but we are exquisitely aware of faces, and photographs of the children's faces revealed something terribly wrong.

The case series were not convincing on their own. As Carrie Randall wrote,

> Despite a shower of case reports, the implication of alcohol as a teratogen in humans was met with skepticism by the medical community. Alcohol was being used at the time to prevent premature labor . . . , so it was difficult to accept the proposal that it could cause harm to the developing fetus. Furthermore, because alcohol was so widely used, it was reasoned that if a causal relationship between prenatal alcohol exposure and birth defects existed, it would have been recognized and reported long before 1973.[23]

These case series were followed by studies with controls.[24] For instance, Beatrice Larroque and colleagues interviewed mothers at a French maternity hospital concerning their alcohol use during pregnancy, then examined their children when they were about 4.5 years old.[25] They compared moderately low and moderately high levels of alcohol consumption, finding that children whose mothers consumed the equivalent of four or more glasses of wine per day had substantially lower performance in a variety of cognitive

assessments. There was also evidence that mothers who drank more alcohol were different from those who drank less: the heavier drinkers had less education, were older, and were more often cigarette smokers; that is, there was evidence of bias in the comparisons, perhaps quite consequential bias. Larroque and colleagues made efforts to remove these biases analytically, in the spirit of the stratification in Table 5.1, but employed other methods. Many subsequent studies also found substantial cognitive deficits among children with prenatal exposures to alcohol, but not all studies concurred.[26] It is, however, no small concern that mothers who drink heavily during pregnancy may differ from those who abstain, differing perhaps in ways that have not been recorded and thus differing in ways that cannot be controlled by analytical adjustments.

Experiments were conducted with laboratory animals. Pregnant mice were given two doses of alcohol by injection into the blood, and later fetuses were found to exhibit facial malformations eerily similar to those found in human children exposed to high levels of alcohol.[27] When seven-day-old rats were given two doses of either a saline solution or alcohol, the brains of rats treated with alcohol "revealed a very dense and widely distributed pattern of neurodegeneration" that "deletes large numbers of neurons from several major regions of the developing brain."[28] Rhesus monkeys given prenatal exposures to alcohol exhibited cognitive deficits when compared with controls.[29] There are, of course, hazards in extrapolating from experiments on animals to effects on humans, particularly in the area of cognition.[30]

Studies based on autopsies and neuroimaging documented structural abnormalities in the brains of humans with substantial prenatal exposures to alcohol.[31] Social or economic differences might induce a spurious association between drinking while pregnant and a child's performance on cognitive tests; for instance, a parent's education and income predict a child's performance on cognitive tests.[32] Are social and economic differences plausible explanations of abnormalities in the structure of a child's brain?

In 2000, the U.S. National Institute on Alcohol Abuse and Alcoholism's *Tenth Special Report to the US Congress on Alcohol and Health* concluded:

Fetal Alcohol Syndrome (FAS) is considered the most common nonhereditary cause of mental retardation. In addition to deficits in general intellectual functioning, individuals with FAS often demonstrate difficulties with

learning, memory, attention, and problem solving as well as problems with mental health and social interactions . . . Estimates of FAS prevalence vary from 0.5 to 3 per 1,000 live births in most populations, with much higher rates in some communities.[33]

Many, perhaps all, of the individual strands of evidence that support the quoted conclusion are vulnerable to alternative interpretations. The woven tapestry of evidence—or the extensively filled in crossword puzzle—is considerably less vulnerable to alternative explanations. Wittgenstein taught us to ask, "What would a mistake here be like?"[34]

Some questions remain about the effects of light alcohol consumption, but there is no compelling evidence that low doses are safe.[35] In 2016, the U.S. Centers for Disease Control and Prevention wrote, "Why take the risk? . . . About half of all US pregnancies are unplanned and, even if planned, most women do not know they are pregnant until they are 4–6 weeks into the pregnancy . . . Sexually active women who stop using birth control should stop drinking alcohol."[36]

Many strands of evidence meet appropriately in support of the claim that sustained and heavy prenatal exposures to alcohol do enormous harm to the developing fetus. Could it be that each strand of evidence is vulnerable to an alternative explanation, as most if not all are vulnerable, and in each case the alternative explanation is correct? It is a logical possibility. Could this be true no matter how many different types of evidence are assembled so long as each one is open to some alternative interpretation? To claim this entails no error in logic: you make no error in logic—you do not contradict yourself—by making this claim. Logical possibility remains. Proof, in the mathematical sense of proof, is lacking. So long as all of the possible alternatives are jointly possible, there is no logical proof. But how important here is the absence of logical proof? In a more general context, Thomas Nagel wrote,

There is no alternative to considering the alternatives and judging their relative merits . . . To dislodge a belief requires argument, and the argument has to show that some incompatible alternative is at least as plausible . . . Someone who said at every point that the apparently law-confirming experimental results were just coincidence would be crazy, but he would not be contradicting himself.[37]

In discussing fetal alcohol syndrome, I have been contrasting the weakness of individual studies that compose a body of evidence no longer weak.

Consensus and Repetition Do Not Make Weak Evidence Strong

There is a scientific consensus concerning fetal alcohol syndrome. How important is it that there is consensus? In the case of fetal alcohol syndrome, the current consensus reflects a change from a previous consensus. The consensus had been that alcohol was not a major focus of concern during pregnancy; then, the consensus changed.

The mere existence of consensus is not a useful guide. We should ask, Does a consensus have its origins and its ground in a rational and comprehensive appraisal of substantial evidence? Has the available evidence been open to vigorous challenge, and has it met each challenge? If a consensus has these origins, then the existence of consensus is of little consequence beyond its important origins. Conversely, a consensus that lacks these origins is of little consequence precisely because it lacks these origins.[38] Knowing the current consensus is helpful in forecasting a vote; having substantial evidence is helpful in judging what is true. That something is standardly believed or assumed is not, by itself, a reason to believe or assume it. Error and confusion are standard conditions of the human mind.

If a design for an observational study produces evidence that is open to a specific alternative interpretation besides a treatment effect, then repeating the same design in many studies will do little or nothing to adjudicate between a treatment effect and the alternative interpretation.[39] Mervyn Susser argued that evidence is strengthened when "diverse approaches produce similar results," particularly when these diverse approaches suffer from diverse weaknesses that are unlikely to align to produce similar results in the absence of an actual treatment effect.[40] For instance, in the case of fetal alcohol syndrome virtually all observational studies of pregnant women face the constant concern that women who drink heavily during pregnancy may differ as mothers from pregnant women who abstain from alcohol. This genuine concern is not an issue in the many controlled experiments on animals, which face legitimate but different concerns. Studies of the neurotoxicity of alcohol in rats suggest a biological

mechanism through which alcohol may produce effects, but leave open the consequences for cognition; however, this legitimate concern does not invalidate studies of cognitive deficits in rhesus monkeys given prenatal exposures to alcohol.

George Polya argues that similar considerations apply in heuristic reasoning in mathematics, developing various qualitative consequences of probabilistic reasoning. Discussing the heuristic reasoning that precedes the proof of a purported theorem, he wrote,

> On the one hand, the examination of a new consequence [of the purported theorem] supplies strong inductive evidence when this consequence has not been made plausible by the consequences examined previously. In practice, this will be the case when this consequence has no immediate relation with the old ones, when it is removed from the preceding, when this new consequence is not only new, but of a new kind. On the other hand, the examination of a new consequence introduces strong inductive evidence when it has a good chance of compromising the theorem. In practice, this will be the case when the examination touches upon a new aspect of the theorem, an aspect of the theorem which had not been previously considered.[41]

✳ Evidence Factors

✳ What Are Evidence Factors?

Typically, if you analyze one data set several times in slightly different ways, you simply repeat yourself. Moreover, if you understand your previous analyses as confirmations of your current analysis, then you are suffering from a statistical version of schizophrenia; the concurring voices you hear are your own, a consensus of one. Indeed, to analyze one data set several times can be dishonest if you report only some of these analyses, especially if you select the reported analyses because only these analyses favor a particular conclusion you wish to reach.

There is an exception, however. The exception is a single planned analysis of an elaborate theory in which several statistically independent tests are performed using the same data, where the different tests are vulnerable to different biases. It takes quite a bit of care in research design to produce this

situation, and in this book I can only sketch the idea, so one must turn to the references for specifics.[42]

To say that two tests are statistically independent under the null hypothesis H_0 of no treatment effect is to say that if H_0 were true then the result of the first test would provide precisely no information about the result of the second. That is something rare, something requiring care in research design, because typically two analyses of the same data strongly predict one another. Because of this statistical independence, the two analyses may be combined as though they came from different studies by unrelated investigators, using methods similar to those used in meta-analyses that combine findings from unrelated studies.

In the simplest situation, to say that the two analyses are vulnerable to different biases is to say that the first analysis would be correct if treatment assignment were ignorable in one sense, and that the second analysis would be correct if treatment assignment were ignorable in a second sense, but massive failures of the first sense would not bias the second analysis, and massive failures in the second sense would not bias the first analysis. This is a much stronger property than statistical independence. One can always pointlessly produce two statistically independent analyses of one data set by splitting the data in half at random and analyzing the two parts separately, but parallel analyses of the two halves will be affected in the same way by the same biases.

✻ An Example of Evidence Factors

To illustrate evidence factors, return to the study by Morton and colleagues of the effects on children of a father's exposure to lead at work. Figure 7.3 reorganizes the data in Figure 7.2. Panel (a) of Figure 7.2 described the lead level in the blood of exposed and control children separately, and panel (a) of Figure 7.3 described the matched pair difference in lead levels, exposed-minus-control or equivalently treated-minus-control. These differences tend to be positive in Figure 7.3, so most exposed children had higher levels of lead than their matched controls. Panels (c) and (d) of Figure 7.2 describe the level of lead for individual children grouped by the level of exposure to lead of the exposed father, and panel (b) of Figure 7.3 groups the treated-minus-control pair differences by the level of exposure of the exposed father.
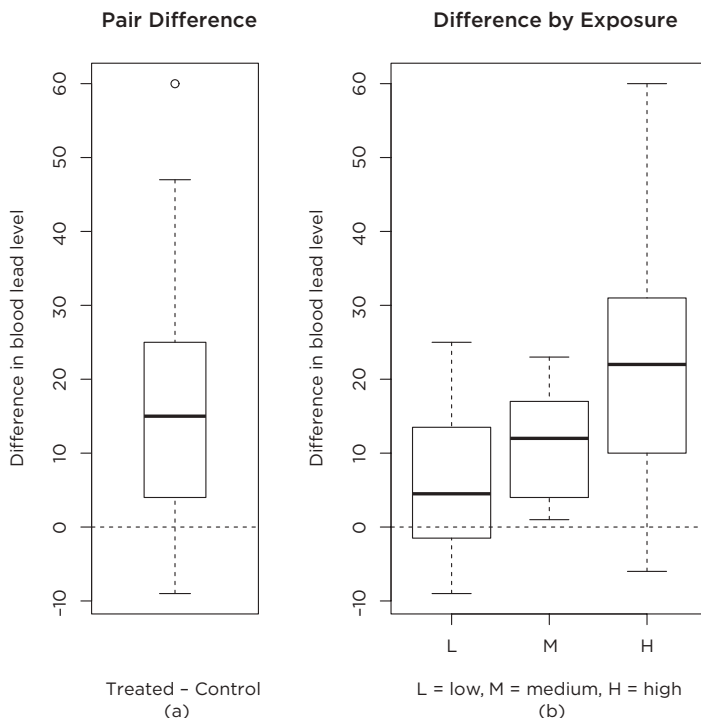
Figure 7.3.   Matched pair differences, treated-minus-control, in levels of lead in children's blood, μg/dl. In each figure there is a horizontal line at zero. Panel (a) shows the differences, while panel (b) separates the differences into three groups based on the level of exposure to lead of the exposed father.

Due to one missing blood reading for one control child, Figure 7.3 refers to 33 pairs of two children.

One might hope that panel (a) of Figure 7.3 is analogous to a simple randomized experiment in which one child in each of 33 matched pairs was picked at random for exposure. One might hope that panel (b) of Figure 7.3 is analogous to a different simple randomized experiment in which levels of exposure were assigned to pairs at random. One might hope that panels (a) and (b) are jointly analogous to a randomized experiment in which both randomizations were done, within and among pairs. All three of these hopes may fail to be realized: there might be bias in treatment assignment within pairs or bias in assignment of levels of exposure to pairs. It turns out, however, that these several hopes are very different hopes; perhaps you will get

part, not all, of what you want. The hopes may be demarcated from each other, so we may ask about the strength of the evidence against the null hypothesis H_0 of no effect if one or another of the various hopes fails to be realized.

Let us consider a simple analysis of a conventional sort, then discuss some properties of the analysis. The analysis tests the null hypothesis H_0 of no treatment effect in two ways, one focused on panel (a) of Figure 7.3, the other on panel (b).

A simple analysis of panel (a) of Figure 7.3 asks whether the pair differences are too often and too substantially tilted toward positive values to be explained by unlucky but fair coin flips in assigning treatments within pairs. The test uses a familiar statistic proposed in 1945 by Frank Wilcoxon, called Wilcoxon's signed rank statistic.[43] It computes the absolute value of each pair difference, ranks those absolute values from one to the number of pairs—here, 33—then sums the ranks for pairs with a positive difference. The statistic turns out to be 499 in panel (a), whereas the largest it could have been is $1 + 2 + \cdots + 33 = 561$; so, as suggested by the appearance of Figure 7.3(a), the differences are heavily tilted toward positive values. If H_0 were true and if treatments were randomized within pairs, then an argument similar to that in Chapter 3 yields the distribution of Wilcoxon's statistic, from which we obtain the two-sided P-value of 1.15×10^{-5}. A boxplot such as panel (a), a boxplot so tilted toward positive values, would have been very improbable in a paired randomized experiment with no treatment effect.

A simple analysis of panel (b) of Figure 7.3 asks whether large pair differences are too often found with higher exposures to be explained by an unlucky but random assignment of exposure levels to pairs. The analysis uses another familiar statistic, Kendall's rank correlation test, which asks whether high values of two variables co-occur too often to be attributed to chance.[44] If H_0 were true and if levels of exposure were randomized among pairs, an argument similar to that in Chapter 3 yields the distribution of Kendall's statistic, from which we obtain a two sided P-value of 0.0104. If levels of exposure had been randomly allocated among pairs in the absence of a treatment effect, it is very unlikely that the steady increase in lead differences with increased levels of exposure in panel (b) of Figure 7.3 would have occurred.

So far, these are two simple statistical analyses of a conventional sort. However, the two analyses stand in an important relationship to one another

when the null hypothesis H_0 of no effect is true. If the assignment of exposure levels were seriously biased and far from randomized, this would invalidate the second test based on Kendall's statistic, but it would do no harm to the first test based on Wilcoxon's statistic, provided that treatments are assigned at random within pairs. In parallel, if the assignment of treatments within pairs were seriously biased and far from randomized, this would invalidate the Wilcoxon test, but it would do no harm to the second test based on Kendall's statistic, provided that exposure levels were assigned at random among pairs. In this sense, the evidence provided by each test is demarcated from the evidence provided by the other test. Moreover, if both assignments were independently randomized, then the two tests would be statistically independent under H_0; that is, if there were no treatment effect, each of the two P-values would provide precisely no information about the other.[45] Because of this, the two P-values can be combined using methods for combining independent P-values; for instance, using Fisher's method yields a combined P-value of 2.04×10^{-6}, considerably smaller than either of its components.[46]

In other words, the two comparisons in Figure 7.3, panels (a) and (b), are each fallible, but in very different ways; discard either one fearing that it is biased, and there is still evidence from the other. Moreover, the two comparisons together provide stronger evidence than either one on its own.

Wilcoxon's statistic and Kendall's statistic are exactly independent in Figure 7.3 under H_0 if both treatment assignments are randomized. This exact independence requires the use of these or similar rank statistics. Many other statistics yield a form of approximate independence in situations like Figure 7.3.[47]

So far, we have considered two extreme possibilities: a comparison, say, the treatment—control comparison in Figure 7.3, panel (a), is either randomized or so severely biased that it is useless. We have seen that either of the two comparisons in Figure 7.3, panels (a) and (b) may be useless without invalidating the other. Must we focus on extreme possibilities? Using the tools in Chapter 9, we may consider less extreme, intermediate cases, in which one comparison is biased to a limited degree but not entirely useless. Considerations of this kind might show that enormous biases affecting either comparison in Figure 7.3 and moderately large biases affecting the other would be insufficient to explain Figure 7.3 as something other than an effect caused by lead from the factory.[48]

There are forms of compulsive checking that do not check anything. Closely parallel analyses of the same data rarely check anything important. In contrast, in seeking evidence factors we are seeking analyses that are not in the slightest bit redundant; indeed, the findings of one analysis provide precisely no information about the findings in the other, so independent confirmation is possible. Wittgenstein remarked about the man who bought "several copies of the morning paper to assure himself that what it said was true."[49] We wish to avoid that man's misunderstanding.

Taking Stock

An elaborate causal theory is one that "predict[s] more and [can] be put more thoroughly to the test."[50] In the limited structure of Chapter 5, ignorable treatment assignment is needed for causal inference but it is not testable; however, an elaborate theory may make it testable. An elaboration of a causal theory is most useful if its predictions help to discriminate between ignorable treatment assignment and a particularly plausible pattern of unobserved biases.

Quasi-experimental Devices

> How can one test a *ceteris paribus* clause severely? By assuming that there *are* other influencing factors, by specifying such factors, and by testing these specific assumptions. If many of them are refuted, the *ceteris paribus* clause will be regarded as well-corroborated.
>
> —IMRE LAKATOS[1]

What Are Quasi-experimental Devices?

Eliminating or Reducing Particular Sources of Ambiguity in Observational Studies

In the absence of random assignment, there are two possible explanations of a difference in outcomes in treated and control groups: a treatment effect or a bias in the formation of the groups. In this sense, an association between treatment received and outcome exhibited is ambiguous. Quasi-experimental devices attempt to reduce an ambiguity in data by looking at more data, to render an ambiguous association less ambiguous by looking at additional associations. Quasi-experimental devices were first considered systematically in 1957 by Donald Campbell.[2]

The Connecticut Crackdown on Speeding

A study by Donald Campbell and H. Lawrence Ross concerned the effects on traffic mortality of a crackdown on speeding in the state of Connecticut

announced on December 23, 1955.[3] Anyone convicted of speeding had their license suspended for 30 days for a first offence, longer for subsequent offences. Did the crackdown save lives?

At the time, Connecticut's governor, Abraham Ribicoff, claimed it did. The study is attractive and instructive because it begins with the data cited by Governor Ribicoff, namely, a modest decline in traffic mortality rates from before to after the crackdown, from 1955 to 1956. Campbell and Ross then discuss how that comparison is relevant but ambiguous in several senses. They add several more years of data for Connecticut, from 1951 to 1959, observing that this comparison is still ambiguous, but less so. In particular, we see that the mortality rate was unusually high in 1955, the year Governor Ribicoff felt he had to take action, and the mortality rate was not low but merely closer to the typical rate in 1956.

When you change policy in reaction to an exceptionally bad number, you need to be cautious about claiming credit for your policy when the number merely goes back to normal. Perhaps you reacted to noise; perhaps the number would have returned to normal without your reaction; perhaps returning to normal is what things normally do. The phenomenon of introducing an ineffective treatment in reaction to a high number that subsequently reverts to a typical number is sometimes called "regression to the mean."

Campbell and Ross then add parallel series of rates for nearby states that did not crackdown on speeding: Massachusetts, New Jersey, New York, and Rhode Island. The patterns in different states are quite variable, and it is debatable whether the pattern in Connecticut is distinctive. For instance, in 1956, after the crackdown, the mortality rate in Connecticut was below that in New Jersey and New York, but it was above that in Massachusetts and Rhode Island. In this one study, seeing more data—more years, more states—made the one-year decline from 1955 to 1956 in Connecticut seem much less compelling as evidence of an effect of the crackdown. Had the additional data turned out differently, the additions might have strengthened the evidence of an effect. A comparison of Connecticut before and after treatment is much more ambiguous than a comparison that adds information easily obtained.

The study by Campbell and Ross of a crackdown on speeding illustrates several themes in quasi-experimentation. A key question is whether treated individuals and controls—here, the several states—were similar before treatment in relevant ways. In attempting to answer this question, particularly the question of relevance, it can be helpful to have a pretreatment measure

of the outcome—here, the mortality rates in the several states in 1955, be-fore the crackdown in Connecticut. Was the outcome, traffic mortality, sim-ilar in treated and control groups before the start of treatment? If not, that would be strong evidence of a biased comparison. Better still would be to observe the outcome measure at several times before treatment—here, the mortality rates between 1951 and 1955 in the several states—to see whether trends in mortality were similar before treatment. Two states could be similar at one point in time, say 1955, yet have very different trends over time—two lines can intersect in 1955 yet be different lines; seeing the states for several years before treatment helps to recognize similar or diverging trends over time.

The treatment occurred abruptly at the end of 1955, so effects should ap-pear subsequent to 1955, but many natural trends that are not treatment effects are gradual, smooth, without a jump at 1955. Seeing treated and con-trol units for several years before and after treatment helps to distinguish treatment effects from gradual trends unrelated to treatment.

Quasi-experimental Devices Help to Avoid Common Cognitive Errors

> A reason can be assigned the wrong weight, as it attracts one's assent too much, or too little.
>
> —ERNEST SOSA[4]

Alfred Mele argued that self-deception is not like lying to yourself. The liar knows that a proposition p is false and intentionally misleads someone else to believe that p is true. Can one person really play both roles, the liar and the believer? Mele considers a person who does not know whether p is true or not-p is true. In logic, not-p is written $\sim p$. Mele exhibited various paths that this person might travel that end in a false belief that p is true. A person might intentionally choose to travel these paths never intending to acquire a false belief about p. In other words, Mele described self-deception as an active cognitive failure, rather than as a lie told to oneself. Two of these paths are "selective focusing / attending" and "selective evidence gathering." Mele wrote:

Selective focusing/attending. Our desiring that *p* may lead us both to fail to focus attention on evidence that counts against *p* and to focus instead on evidence suggestive of *p*.

Selective evidence gathering. Our desiring that *p* may lead us to overlook easily obtainable evidence for ~*p* and to find evidence for *p* that is much less accessible.

Desires have effects on the vividness and availability of data, which in turn have effects on belief acquisition, even though the believer does not try to produce any of these effects.[5]

Quasi-experimental devices discourage these and other cognitive errors. The devices focus attention on aspects of the data in hand that might reveal unmeasured biases if biases are present, aspects that might distinguish an actual treatment effect from a particular unmeasured bias. The devices also call attention to additional data that might readily be obtained. A successful quasi-experiment feels like what it is intended to be: a comprehensive and fair-minded weighing of alternative interpretations in light of each available source of relevant evidence.

This chapter discusses two of the many quasi-experimental devices: multiple control groups and counterparts. Other devices are mentioned briefly.

Multiple Control Groups

What Are Multiple Control Groups?

Often there is a reason, or various reasons, that one person receives treatment and another does not. What seems to be an effect caused by the treatment may instead reflect the reasons people end up as treated or untreated. If people are denied treatment for several reasons, it can be informative to form several control groups, each denied treatment for a different reason. Do people who were denied treatment for different reasons have similar outcomes? Or do outcomes vary widely among control groups that were denied treatment for different reasons? Substantial differences in outcomes among control groups cannot be effects caused by the treatment and are likely to indicate some bias in the way treatments were assigned. If outcomes are

similar in several control groups and quite different in the treated group, then this pattern is, at least, not inconsistent with an effect caused by the treatment.

To be useful, two control groups must differ in some consequential way. If having two control groups were valuable merely because there are two, rather than one, then any one control group could be split in half at random to form two control groups, but of course this would provide no information about unmeasured biases—whatever is wrong with one control group is wrong with the other as well.

The study in Chapter 6 of the Dutch famine by Zena Stein, Mervyn Susser, Gerhart Saenger, and Francis Marolla is one example: some controls escaped prenatal exposure to the famine by virtue of where they lived, others by virtue of when they were born.[6] Each control group is imperfect. Are their imperfections consequential? If we saw that these several control groups had very different cognitive outcomes at age 18, then we would have serious concerns about their role as control groups. If we saw that different control groups have similar outcomes, we would strengthen the claim that the reasons the control groups differ—here, time and location—do not explain differences in cognitive outcomes between the treated group and all of the control groups. A strengthened claim is not an invincible claim. Still, stronger is better.

Multiple control groups are a commonly used quasi-experimental device. Illustrations follow.

Multiple Control Groups in a Study of the Possible Risks of Antibiotics

Wayne Ray and colleagues asked whether a particular antibiotic, azithromycin, increases the risk of death from cardiovascular causes.[7] They compared people during periods of time when they received azithromycin to two control groups. One control group consisted of people during periods of time when they were not receiving any antibiotic medication. The other control group consisted of people during periods of time when they were receiving a different antibiotic, specifically amoxicillin. Each control group has a limitation. If azithromycin and amoxicillin both increased the risk of

cardiac death, then this could easily be missed by comparing people receiving either azithromycin or amoxicillin. On the other hand, a person receiving azithromycin has an infection requiring treatment, whereas a control not receiving an antibiotic probably does not have such an infection, so a comparison of azithromycin to an untreated control risks confusing the harm caused by the infection with harm caused by the medication used to treat the infection.

After adjusting for observed covariates x_i, Ray and colleagues found a small increase in risk of cardiovascular death among people receiving azithromycin when compared with either control group. This conclusion is less ambiguous than a conclusion based on either single control group, and less ambiguous is better.

Multiple Control Groups in a Study of the Possible Risks of Inhaled Corticosteroids

An analogous comparison with different conclusions was made by T. P. Van Staa, H. G. M. Leufkens, and C. Cooper in a study of the possibility that inhaled corticosteroids increase the risk of fractures. Inhaled corticosteroids are often used to treat asthma but have been associated with fractures. Do inhaled corticosteroids cause the fractures with which they are associated? A causal effect of inhaled corticosteroids is not implausible because "oral corticosteroid therapy is an established cause of osteoporosis."[8] At the same time, taking a pill is different from inhaling a medication, and both are different from applying a salve to the skin. It would not be surprising if the same or similar medications had different side effects depending upon the dosage and manner of administration. The study used the United Kingdom General Practice Research Database.

Van Staa and colleagues compared users of inhaled corticosteroids to two control groups who did not use inhaled corticosteroids. One group used a bronchodilator and some other form of nonsystematic corticosteroids. The other group used some other form of nonsystematic corticosteroids but without a bronchodilator. In both control groups, about 75% of patients had used a topical corticosteroid. Van Staa and colleagues expected nonsystematic corticosteroids, such as topical corticosteroids, to have no effect on the

risk of fractures; rather, the controls were required to have exposure to non-systematic corticosteroids to demonstrate that they were active participants in the U.K. General Practice Research Database.[9]

Van Staa and colleagues found an elevated rate of fractures among users of inhaled corticosteroids when compared with the second control group, but not when compared with the first control group who used a bronchodilator. They interpreted this as evidence that inhaled corticosteroids are not the cause of the factures. Specifically, they raised the possibility that users of both inhaled corticosteroids and bronchodilators were more severely ill, and hence perhaps less physically active, than the controls who used just other forms of nonsystematic corticosteroids, and that the association with fractures was a byproduct of physical inactivity.

In this example, the use of two control groups signaled that the treatment is less likely, perhaps unlikely, to be the cause of an outcome with which it is genuinely associated. The situation would be less clear—more ambiguous—if either control group had been used alone.

A finding that a treatment is not the cause of a genuine association can be just as important as a finding that it is the cause. It is just as important to know that a treatment does not have harmful side effects as to know that it does have such effects.

Multiple Control Groups in a Study of Reimbursement for Mental Health Services

Health insurance plans in the United States are often less generous in paying for mental health or substance abuse services than for medical services. For instance, there may be higher deductibles and copayments for mental health services or other limitations on the use of such services. Presumably these limitations are intended to save money for the insurer, either by transferring some costs to the patient or by discouraging extensive use of mental health services. Even if one regarded saving money in this way as the desired goal, it is unclear whether this tactic actually saves money. It is possible that the tactic discourages inexpensive treatment of manageable problems; then these untreated problems become larger, requiring expensive treatment. For example, perhaps a patient who regularly takes an inexpensive antipsychotic medication may thereby avoid an expensive psychiatric hospitalization.

In 1999, President Bill Clinton ordered a change in this policy for employees of the U.S. government covered under the Federal Employees Health Benefits Program. The change took effect in January 2001 and required parity, meaning parallel coverage for mental health and medical services. What are the effects of parity?

There have been several evaluations of the effects of this change in policy. The evaluation by Frank Yoon, Haiden Huskamp, Alisa Busch, and Sharon-Lise Normand compared a treated group consisting of a sample of federal employees in 2001 to two control groups.[10] One control group consisted of a nonoverlapping sample of federal employees in 2000 before the program went into effect, a similar group of people but in a different year. The other control group consisted of employees at private companies covered by unaffected private health insurance plans in 2001, a different group of people studied in the same year. Each control group has a small imperfection. Are the imperfections consequential?

Sensibly, the study was designed so that each of the three groups had a full year of baseline measures or observed covariates, x_i, in either 1999 or 2000. Using these baseline measures, Yoon and colleagues compared people whose mental health appeared similar in the year before the comparison year. Their analysis focused on people who had, in the baseline period, a severe diagnosis such as schizophrenia; then the comparison was of service utilization in the subsequent comparison year, either 2000 or 2001. Specifically, they matched people for age, gender, mental health expenditures in the baseline year, employee or dependent, and many specific diagnostic categories.

Yoon and colleagues found evidence that the treated group was more likely than both controls to use at least one mental health service, was more likely to receive a mental health prescription, but was substantially less likely to use inpatient mental health services. So there were indications that the treated group received more inexpensive routine care, and less expensive or emergent care.

The analysis by Yoon and colleagues is instructive in various ways. In part, they made extensive use of multivariate matching for a year of baseline measures, so the three groups look comparable in measured covariates prior to the comparison of outcomes. Additionally, they used a structured approach to the analysis, rather than simply making lots of disorganized comparisons. Their structured comparison put the main questions first

to maximize the chance of finding something if there was something there to be found, but it went on to demonstrate that the two control groups were nearly equivalent in terms of outcomes. The topic of structured comparisons is discussed in a later starred or optional section of this chapter.

Control by Systematic Variation

What considerations are important in selecting two control groups? In 1969, Donald Campbell argued for "control by systematic variation."[11] To explain the use of "control by systematic variation" in observational studies of people, Campbell quoted from M. E. Bitterman, a psychologist who compared learning in different species, such as fish and rats:

> I do not, of course, know how to arrange a set of conditions for the fish which will make sensory and motor demands exactly equal to those which are made upon the rat in some given experimental situation. Nor do I know how to equate drive level or reward value in the two animals. Fortunately, however, meaningful comparisons still are possible, because for *control by equation* we may substitute what I call *control by systematic variation*. Consider, for example, the hypothesis that the [ostensible difference in learning between fish and rats] . . . is due to a difference, not in learning, but in degree of hunger. The hypothesis implies that there is a level of hunger at which the fish will show progressive improvement, and put this way, the hypothesis becomes easy to test. We have only to vary level of hunger widely in different groups of fish, which we know well how to do. If, despite the widest possible variation in hunger, progressive improvement fails to appear in the fish, we may reject the hunger hypothesis.[12]

Instead of trying to make the rat and the fish equally hungry, Bitterman shows that the fish does not learn what the rat learns no matter how hungry the fish becomes. The level of hunger is not measured, but it is widely varied, and wide variations in hunger do not alter the outcome, the degree of learning, so inequality in hunger is not a plausible explanation of the ostensible difference in learning. The attraction here is that control by systematic variation uses something that is observed to render

implausible an alternative explanation that involves something that was not observed.

In observational studies of people, control by systematic variation can sometimes resolve a claim about an unobserved covariate u_i without measuring it. One finds two control groups that both resemble the treated group in terms of measured covariates, x_i, but are known to be very different from each other in terms of u_i, even though u_i is not observed for any person. If control groups that are very different in terms of u_i have similar outcomes, and these outcomes are very different from the treated group, then treated-versus-control differences in u_i are not a plausible explanation of the ostensible effect of the treatment.[13]

One might think that it would be difficult to find control groups that systematically vary an unobserved covariate u_i, but Campbell denies this: "Controlling plausible rival hypotheses through supplementary variation . . . [is] a very general technique of partial or inferential control useable for many settings in which direct or complete control is not possible . . . Clarity of inference sometimes may be improved by deliberately reducing the quality of part of the data."[14] What does he mean by "deliberately reducing the quality of the data"? If the treatment was introduced at midnight on Tuesday, perhaps the best comparison is of treated-Wednesday versus untreated-Tuesday because trends over time that might confound the treatment effect are minimized; however, an additional comparison with Monday—an inferior control group further away in time—may help to show that no strong trend was present prior to the start of treatment. If the treated group is highly motivated, then a highly motivated control group is best, but an additional comparison with an unmotivated control group may help to show that it is the treatment, not the degree of motivation, that is driving the outcomes. That a good but imperfect control group and a somewhat defective control group have similar outcomes may demonstrate that the specific defect in question is not of great importance in predicting outcomes, perhaps strengthening the case that the good control group is good enough.

Several different recent strategies use the computer to build two control groups with systematic variation in mind.[15] These techniques use controls that might otherwise be discarded because they are different from treated individuals in terms of some measured covariates.

✱   Organizing Analysis with Two Control Groups

Use of quasi-experimental devices, such as two control groups, often entails making several simple empirical comparisons. An organized, planned approach to such comparisons is better than making a comparison here, another there, and then pondering how to write it all up in a convincing way. Too often, disorganized comparisons look like "fishing expeditions," a term scientists use to describe the search for a publishable finding, rather than for a true and truthful finding.

Informal discussions of quasi-experimental devices imagine and hope for sharply defined patterns. For instance, one might hope to see much higher responses in the treated group, and two control groups with lower identical responses; that pattern might support a claimed treatment effect. Alternatively, one might hope to see responses in the treated group that are identical to those in one of the control groups and very different from the other control group; that pattern might support the claim that a genuine association of an outcome with treatment is not an effect caused by the treatment. Commonly, the investigator does not see situations that are so sharply defined. In part, data are noisy—after all, the different groups contain different people—so actual groups are never identical. In part, besides noise, the underlying situation may be more complicated than the investigator hopes.

In disorganized analyses with two control groups, three mistakes are common. First, the investigator may hope to show that the two control groups had similar outcomes, but it is a mistake to try to show this by testing whether the outcomes are different in the two control groups. If you test the hypothesis of no difference in outcomes in two control groups, and if you reject that hypothesis with a P-value less than α, conventionally $\alpha = 0.05$, then you have provided evidence that at least one of the control groups would yield biased estimates of the treatment effect when compared with the treated group.[16] That may be a useful finding. However, if you fail to reject in this test, then you have not provided evidence that the two control groups are the same or similar. Failure to reject a hypothesis means failing to provide evidence that the hypothesis is false; however, it does not mean providing evidence that the hypothesis is true. Perhaps you did not provide much evidence one way or the other; see the discussion of this mistake in Chapter 3. To show two groups are similar, we need to reject every hypothesis that says they are substantially

different; that is, we need either an equivalence test or a three-sided test.[17] A three-sided test looks for both equivalence and difference, and it is appropriate and efficient when comparing control groups to each other.

The second mistake concerns testing several hypotheses, the so-called problem of multiplicity. If you test one hypothesis at level $\alpha = 0.05$, then you run a 1-in-20 or 5% chance of falsely rejecting a true hypothesis—that is, of getting a P-value of 0.05 or less when the hypothesis you reject as false is actually true. Perhaps a 1-in-20 risk is tolerable. However, if you test 20 true hypotheses at level $\alpha = 0.05$, then you expect to falsely reject one of them, $1 = 20 \times 0.05$; that is, you expect to get a P-value of 0.05 or less for some hypothesis that is actually true. Stated informally, with 20 tests you expect to publish a false finding if you do nothing to control the problem of multiplicity. This issue, multiplicity, is the reason scientists worry about fishing expeditions: test enough hypotheses and you will get plenty of small P-values, even if all the hypotheses are true, just because you tested so many hypotheses. One approach that addresses multiplicity imposes a tougher standard; for instance, the common use of the Bonferroni inequality would require the smallest P-value to be at most $0.05 / 20 = 0.0025$ before any hypothesis is rejected when 20 tests are performed.[18] Though familiar, the Bonferroni method is not the best method for structured situations, such as a treated group and two control groups. A better method is described here.

The third mistake is to reduce the chance of finding anything—to reduce the power—by failing to exploit the highly structured nature of a treated group and two control groups. In highly structured problems, multiplicity can be addressed by "quitting," rather than by the Bonferroni approach of requiring very small P-values. Quitting is my informal term for various formal procedures such as "closed testing" and related methods.[19] All tests are done at the same conventional level, commonly $\alpha = 0.05$. The structure leads us to test one hypothesis first. Typically, if this hypothesis is rejected with a P-value of $\alpha = 0.05$ or less, then we test one or more additional hypotheses; however, one or more large P-values may terminate all testing. When designed correctly, a "quitting" strategy may test many hypotheses, each one at level $\alpha = 0.05$, yet the chance of falsely rejecting any true hypothesis is at most 0.05, despite doing many tests. When designed correctly, a quitting strategy works by setting priorities, doing first what is most important, pursuing success, and terminating failure.

One reasonable, organized analysis with two control groups proceeds as follows. It is a quitting strategy. First, the treated group is compared with an average combination of the two control groups. So this first test uses all the controls. If that test produces a P-value greater than $\alpha = 0.05$, then testing stops with no difference discovered. Otherwise, in the second step, the hypothesis of no difference is rejected, and testing continues by comparing the treated group with each control group separately, two tests each performed at the $\alpha = 0.05$ level. That second step might fail to reject anything, might find a difference for one control group but not for the other, or it might find a difference for both control groups. Testing continues to the third step only if both P-values in the second step are $\alpha = 0.05$ or less. The third step, if it is reached, tries to show that the two control groups differ less from each other than either group differs from the treated group, a form of equivalence testing.[20] Key issues are tested first, whereas supporting issues are tested only if the key tests provide reason to be interested in these supporting issues. The goal is to avoid the three common mistakes previously discussed when comparing a treated group with two control groups.

Counterparts

What Are Counterparts?

> There can be no disadvantage in noticing obvious analogies provided that we are prepared to recheck and revalue them . . . *Clarified analogies.* Analogy is often vague. The answer to the question, what is analogous to what, is often ambiguous. The vagueness of analogy need not diminish its interest and usefulness; those cases, however, in which the concept of analogy attains the clarity of logical or mathematical concepts deserve special consideration. Analogy is similarity of relations. The similarity has a clear meaning if the relations are governed by the same laws . . . In general, systems of objects subject to the same fundamental laws (or axioms) may be considered as analogous to each other, and this kind of analogy has a completely clear meaning . . . Isomorphism is a fully clarified analogy.
>
> —GEORGE POLYA[21]

True controls closely resemble treated individuals except for the treatment. Counterparts are not true controls. Counterparts stand in a defined relation to the treated and control groups, differing in a well-defined sense but analogous

in other well-defined senses. If counterparts are governed by the same laws or forces as treated and control groups, we may study the operation of those forces in the absence of treatment by studying the counterparts. In this way, counterparts may indicate whether those forces can or cannot readily explain the difference in outcomes in treated and control groups. Often the goal is to reinforce the flimsy claim, "Either there is a treatment effect, or else certain specific forces made outcomes different in treated and control groups." The reinforced claim says, "Either there is a treatment effect, or else certain specific forces made outcomes different in treated and control groups, but only if those forces acted very differently on treated and control groups than on their analogous counterparts." Reinforcement strengthens a claim; it does not make it unbreakable. Still, stronger is better.[22]

A common use of counterparts occurs when a certain type of person is sure to receive the control before a particular date, and is sure to receive the treatment after a particular date. Such a treated-versus-control comparison may be moderately convincing on its own if we have strong reason to doubt that the world changed in relevant ways from before that particular date to after. Counterparts are added to such a study in the hope of providing just such a reason, or else to instill caution because even the untreated counterparts exhibited substantial change in outcomes with the passage of time. These counterparts were never eligible for treatment because of some way they differed from the treated and control groups, so the counterparts are not true controls. We would not compare the treated group to the untreated counterparts because we can see in the data that they differ in an important way. Nonetheless, we may learn about forces that act on everyone in the same way by studying how those forces affect the counterparts.[23]

Counterparts need not refer to time. Suppose that one hospital uses general anesthesia for knee replacement surgery and another hospital uses regional or local anesthesia, and an investigator compares postoperative cognitive outcomes at these two hospitals for patients undergoing knee surgery. Although use of general anesthesia is directly relevant to postoperative cognitive outcomes, there could be other differences between these two hospitals that are also relevant. If both hospitals use general anesthesia for hip replacement surgery, then patients undergoing hip replacement surgery serve as counterparts. Obviously, they are counterparts and not true controls—they underwent a very different type of surgery—but it would be reassuring to see that cognitive outcomes were similar at the two hospitals for a type of

surgery in which both hospitals used general anesthesia. As is often true with quasi-experimental devices, this illustration directs us to consider data that is readily available but might be ignored.

Injury, Compensation, and Incentives

Social programs with desirable objectives may have unintended consequences, and much of the debate that surrounds such programs weighs evidence about their intended and unintended effects. If such debates are to be more than a repetition of ideological commitments, then the programs need to be studied empirically. An example of the use of counterparts occurs in one such study by Bruce Meyer, Kip Viscusi, and David Durbin.[24] They were interested in whether higher levels of worker compensation for injuries create an incentive for workers to stay out of work for a longer period of time.

Worker compensation programs are constructed to be humane while avoiding inappropriate incentives. Typically, injured workers receive less in compensation than they were earning while working, but enough to get by and recover. Also, there is typically a ceiling so that a highly paid worker—say the company president—does not receive an enormous amount in injury compensation. This ceiling periodically becomes dated as a consequence of inflation so that it no longer affects only highly paid workers but also typical workers, and at this point the ceiling may be raised. An increase in the ceiling only affects workers who had reached the old ceiling, that is, workers with somewhat higher wages.

On July 15, 1980, the state of Kentucky raised the maximum benefit from $131 to $217 per week, an increase of 66%. On January 1, 1982, Michigan increased the maximum benefit from $181 to $307 per week, an increase of 70%. In other words, the ceilings rose substantially, but the increase had consequences only for workers who had reached the old ceiling. Meyer, Viscusi, and Durbin compared treated and control workers whose incomes would have entitled them to the old maximums of $131 in Kentucky and $181 in Michigan in the early time period, and would have entitled them to the new maximums of $217 or $305 in the later time period. Control workers were injured in the period before the increase in compensation, treated workers were injured in the period after the increase. Did treated workers receiving higher compensation stay out of work longer? This is

not an implausible comparison as it stands, but it is open to the objection that economic conditions fluctuate with time; perhaps a difference in time out of work reflects changed economic conditions, not the change in compensation.

To address this concern, Meyer, Viscusi, and Durbin examined counterparts, specifically, injured workers with much lower wages. The increase in the compensation ceiling had no consequence for the compensation received by workers with lower wages—they were not affected by the ceiling before or after it was raised. Obviously, the counterparts are not true controls: they have much lower wages than both the treated and control groups, presumably worked in different jobs, may have been injured at work in very different ways, and may possess very different resources to draw upon in an emergency. Nonetheless, the counterparts faced the same changes in economic conditions faced by treated and control groups, and faced them with no change in their compensation. True, a change in economic conditions could affect lower-wage workers differently from higher-wage workers; that is, counterparts strengthen the comparison but do not make it unbreakable. What possible argument could there be for not looking at evidence that will strengthen the comparison?

A sensible feature of the study by Meyer, Viscusi, and Durbin is that their treated group received the full effect of the increase in the ceiling on compensation, $131 to $217 in Kentucky and $181 to $307 in Michigan, while their control and counterpart groups received nothing. In making this comparison, they excluded from consideration some people who received a small increase in benefits because of a graduated scale for benefits.[25] One of the several senses in which this strategy is sensible is discussed in Chapter 11 in connection with design sensitivity: excluding from consideration people who received trivial doses of a treatment tends to make a study less sensitive to biases from unmeasured covariates.

Table 8.1 describes two matched pair comparisons, one of the male treated and control groups, and the other of their male counterparts.[26] Not shown in Table 8.1, workers in Kentucky were matched to workers in Kentucky, and workers in Michigan were matched to workers in Michigan. Figure 8.1 displays weekly wage prior to injury and age for the treated (HA) and control (HB) groups and their counterparts (LA and LB). Because the treated and control groups were defined by a wage high enough to be affected by the ceiling on compensation, the treated and control groups have much

Table 8.1. Two matched pair comparisons before and after the rise in the ceiling for injury compensation for affected higher-wage workers and unaffected lower-wage counterparts

Covariate	Treatment/control comparison Higher-wage workers		Counterparts, all untreated Lower-wage workers	
	Before = Control	*After = Treated*	*Before*	*After*
Sample size	1,224	1,224	1,286	1,286
Male (%)	100	100	100	100
Age (years, mean)	37	37	31	31
Injury (%)				
Head	3	3	4	4
Upper extremities	25	24	32	32
Trunk	14	13	12	12
Lower back	27	29	24	24
Lower extremities	26	26	24	24
Occupational disease	0	0	0	0
Industry (%)				
Manufacturing worker	18	18	33	32
Construction worker	21	21	13	15
Wage, $ per week (mean)	519	517	210	209
Married (%)	85	85	57	56

higher wages than their counterparts. That difference in wages is built into the design of the study; it is part of the definition of the counterparts. In addition to this defined difference, there are many other differences: the counterparts are younger, are less often married, are less often construction workers and more often manufacturing workers, and have a different pattern of injuries. It is reasonable to compare treated and control groups—at least, they look similar in terms of observed covariates—and it is reasonable to compare counterparts before and after—they look similar in terms of observed covariates; however, by definition, the counterparts look very different from the treated and control groups.

Figure 8.2 shows the outcome, the duration of benefits in weeks on a log scale, where the minimum duration is set to one week. The log scale emphasizes the shorter durations for most workers but lets us see some much longer durations for a few workers, presumably workers suffering from severe injuries. In the first two boxplots, for the control (HB = high, before) and treated (HA = high, after) groups, after the ceiling was increased, the
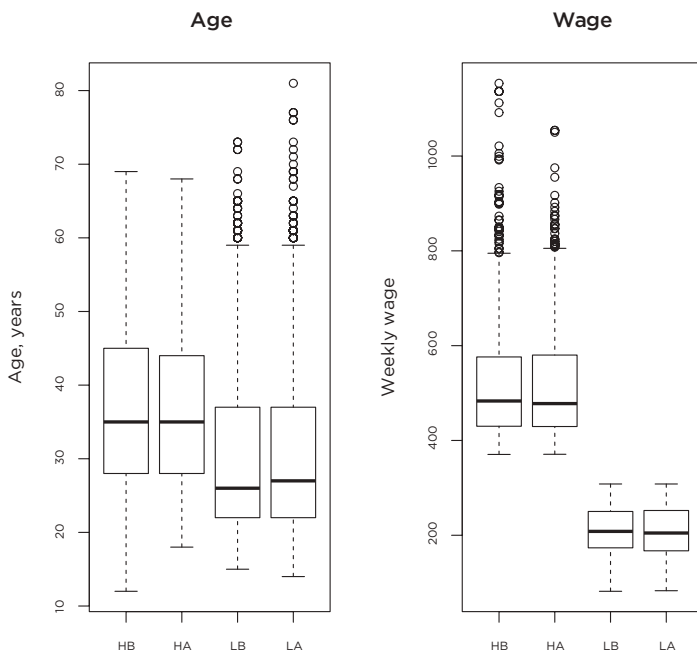
Figure 8.1.   Covariate balance for two covariates, age and prior wage, in the study of the effects of the increase in the ceiling on compensation for injury. The four groups are workers with higher (H) or lower (L) wages before injury, and after (A) or before (B) the increase in the compensation ceiling. Lower-wage workers—the counterparts—are unaffected by the ceiling before and after its increase. The increase in the ceiling raised the compensation for the higher-wage workers in the period after the increase.

median duration of benefits rose by one week from 4 weeks to 5 weeks, the lower quartile remained at 2 weeks, and the upper quartile rose from 8 weeks to 10 weeks. Judged by Wilcoxon's signed rank statistic, discussed in Chapter 7, this change would be improbable had it occurred in a randomized experiment with no treatment effect, the two-sided P-value being 5.1×10^{-7}. In the second two boxplots, for the counterparts (LB = before and LA = low after), the median remained at 4 weeks, the lower quartile rose from 1 week to 2 weeks, and the upper quartile rose from 7 weeks to 8 weeks, with a two-sided P-value of 0.057. There is, then, some indication of longer durations when benefits increased, but the typical magnitudes are not large.
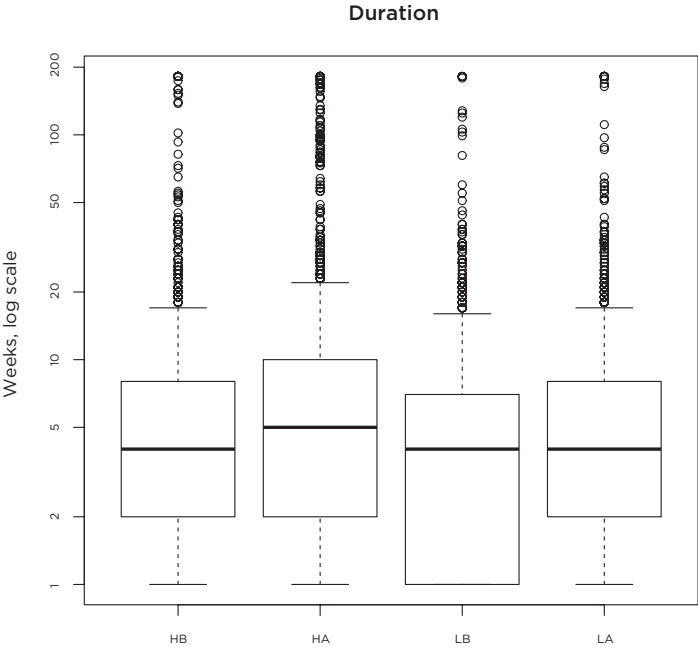
Duration

Figure 8.2. Duration of injury compensation benefits in weeks on a log scale. The minimum is set to one week. The four groups are workers with higher (H) or lower (L) wages before injury, and after (A) or before (B) the increase in the compensation ceiling. Lower-wage workers—the counterparts—are unaffected by the ceiling before and after its increase. The increase in the ceiling raised the compensation for the higher-wage workers in the period after the increase.

The comparison in Figure 8.2 is sometimes called the "method of difference-in-differences," perhaps not the best phrase. The phrase suggests that we ask the following question. Is the difference different? Is the after-minus-before difference different for higher-wage workers than for lower-wage workers? Remember, only the higher-wage workers were affected by the increase in the ceiling on compensation. If the after-minus-before difference was the same for higher- and lower-wage workers, it would seem to indicate that the change in compensation did not produce the difference—rather, perhaps a shift in economic conditions or something else produced the difference.

Figure 8.3 looks at matched pair differences, after-minus-before, for the higher- and lower-wage workers. Positive values in Figure 8.3 indicate that
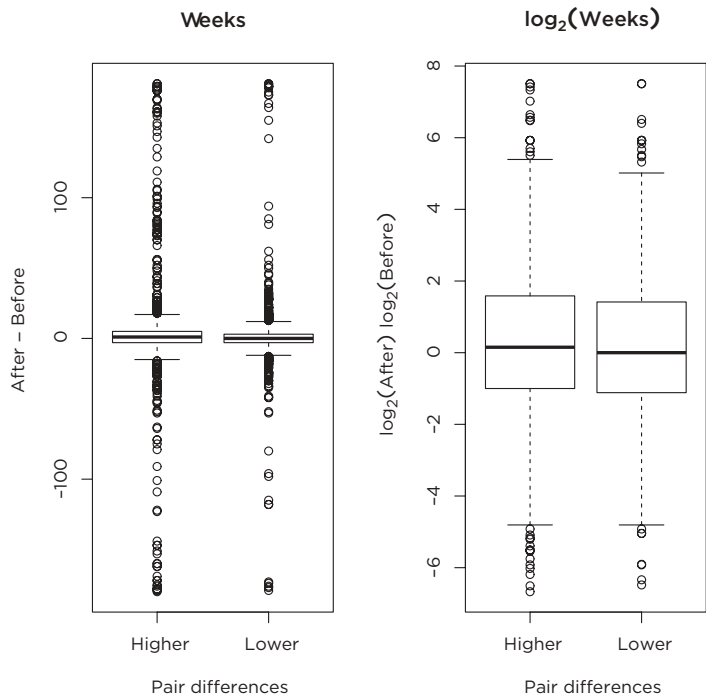
Figure 8.3.   After-minus-before matched pair difference in duration of injury compensation. The left panel is the difference in weeks. The right panel is the difference in $\log_2$(weeks). The higher wage group (Higher) was affected by the change in the ceiling on compensation, whereas their lower wage counterparts (Lower) were not. As throughout, the minimum duration is set to one week.

the worker injured in the after period stayed out of work longer than the matched worker in the before period. The left side of Figure 8.3 looks at the difference in duration in weeks, and on this scale the visual attention focuses on a small number of pair differences in which one worker was out for many weeks, presumably because of a severe injury. The right side of Figure 8.3 takes logs of the duration before computing the differences. On the log scale, we can see what is happening for typical workers.

Figure 8.3 uses base-2 logs, or $\log_2(\cdot)$. Using base 2 makes logs easy to read. Base 2 logs are understood in terms of doublings. A difference of one means one-doubling. For instance, $\log_2(2) - \log_2(1) = 1$ because $2 = 2 \times 1$, but also $\log_2(10) - \log_2(5) = 1$ because $10 = 2 \times 5$, and $\log_2(80) - \log_2(40) = 1$

because $80 = 2 \times 40$; that is, in each case, there is one-doubling. In general, $\log_2(a) - \log_2(b) = 1$ if $a = 2 \times b$. In parallel, a difference of two means two-doublings. For instance, $\log_2(4) - \log_2(1) = 2$ because $4 = 2 \times 2 \times 1$, but also $\log_2(20) - \log_2(5) = 2$ because $20 = 2 \times 2 \times 5$. In general, $\log_2(a) - \log_2(b) = 2$ if $a = 2 \times 2 \times b = 4 \times b$. In complete generality, $\log_2(a) - \log_2(b) = k$ if $a = 2^k \times b$, that is, if b must be doubled k times to produce a. Changing from weeks to $\log_2$(weeks)—changing from the left panel to the right panel of Figure 8.3—had a large effect on the appearance of the boxplots.

What about changing the base of the logarithm? What is the effect on a figure like the right panel of Figure 8.3 of changing the base of the logs—say from natural logs to base 10, or from base 10 to base 2? Changing the base does not change the appearance of the plot; rather, it changes the numbers attached to the axis of the plot.[27] Changing to base 2 logs has just one effect: it makes the numbers on the y-axis easy to read. A difference of one on the y-axis of the right panel of Figure 8.3 means one-doubling.

In the right panel of Figure 8.3, the median after-minus-before change is close to zero for both higher- and lower-wage workers, but it is slightly higher for the higher-wage workers. In other words, the difference-in-differences is slightly positive on a log-scale. Although the difference-in-differences does not look large in the right panel of Figure 8.3, the P-value testing equality is very small, 6.18×10^{-13}, so the difference is not plausibly due to chance, and the estimate from the difference-in-differences is a 17% longer duration.[28] So, once again, there is some indication that higher-wage workers stayed out of work longer after the benefit ceiling was raised, but the difference is not extremely large.

∗ Can You Eliminate Unmeasured Bias by Subtraction?

Counterparts are often useful. This was evident earlier in the Connecticut crackdown on speeding and in the study of injury compensation. Most of the literature using counterparts refers to them as "nonequivalent controls" and does not mention the phrase "difference-in-differences." The phrase "difference-in-differences" suggests that the situation is simpler than it is, that unmeasured biases can be removed by subtraction, by differencing twice. It is not difficult to see that this cannot work in general.
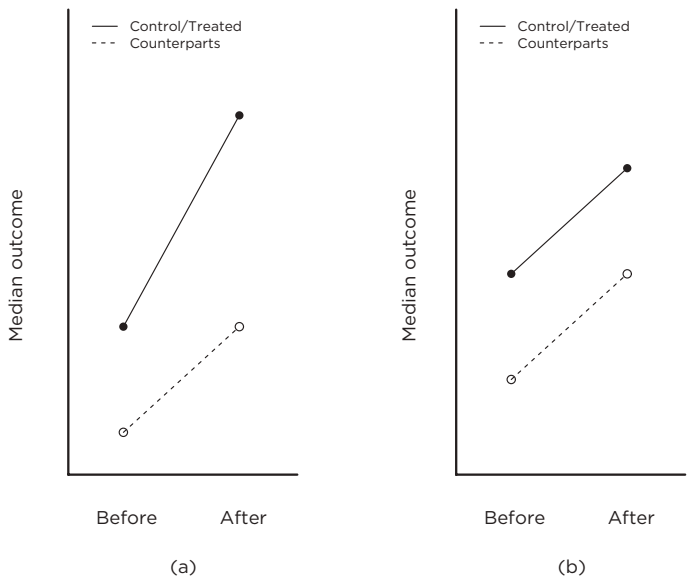
Figure 8.4.   Stylized plots of difference-in-differences. There is a treated group in the after period, a control group in the before period, and untreated counterparts. The dots represent the median response in each group; they correspond with the middle line in the boxplot.

Figure 8.4 is a stylized illustration. That is, Figure 8.4 is an artificial example crafted to illustrate a particular issue. The boxplots in Figure 8.2 are each replaced in Figure 8.4 by a single point, one point per group, representing the median response in the group, the middle lines in the boxplots in Figure 8.2. However, Figure 8.4 is an artificial example. There is a treated group in the after period, a control group in the before period, and untreated counterparts in the before and after periods. The situation is analogous to Figure 8.2.

Panels (a) and (b) of Figure 8.4 represent two possible outcomes for a study. In panel (a), the untreated counterparts were not comparable to the controls in the before period. The counterparts increased from before to after, but the treated group increased by more. That is, the two lines are farther apart in the after period than in the before period. Because of this, the difference-in-differences is a positive number. A mechanical view of counterparts—a

view in which you mechanically compute the difference-in-differences and call it a treatment effect—would say that there is a treatment effect in panel (a).

The situation looks different in panel (b). In panel (b), in the period before treatment, the controls and their counterparts are different, and in the period after treatment, the treated group and their counterparts are different, but the difference before is the same as the difference after. A mechanical view of counterparts would say that there is no treatment effect in panel (b) because the difference-in-differences is zero.

The problem with the mechanical view is that panel (b) in Figure 8.4 was produced from panel (a) by taking logarithms. It is well known that taking logarithms can change the value of the difference-in-differences, and in Figure 8.4 the change is from a positive number to zero.[29] However, it makes no sense, scientifically or mathematically, to say that the null hypothesis of no treatment effect is false for a response, R_i, but true for its logarithm, $\log(R_i)$. If raising the ceiling on injury compensation caused you to stay out of work for additional weeks, then it also caused an increase in the logarithm of the number of weeks you were out of work. The mechanical view of counterparts is simply mistaken. Unmeasured biases cannot be removed in a mechanical way by subtraction. The literature that refers to counterparts as "nonequivalent controls" avoids this mistake and emphasizes that the difference-in-differences may fail, in various ways, to estimate the treatment effect.[30]

Figure 8.5 considers additional stylized plots depicting three possible outcomes of a treatment and control comparison with counterparts. In panel (c) of Figure 8.5 the counterparts and controls were comparable in the before period, the counterparts did not change from before to after, but the treated group had higher responses. Arguably, the pattern in panel (c) comes closest to what is expected from a treatment effect in the absence of visible bias. In panel (d), the controls and their counterparts were comparable in the before period, both groups increased from before to after, but the treated group increased by more than its counterpart. Panel (d) is less satisfactory than panel (c) because the counterparts changed in the absence of treatment. In panel (e), the controls and their counterparts differed in the before period, the counterparts did not change, and the treated subjects in the after period had higher responses than the controls in the before period. Panel (e) is less satisfactory than panel (c) because the counterparts and the controls differed in the absence of treatment.[31]
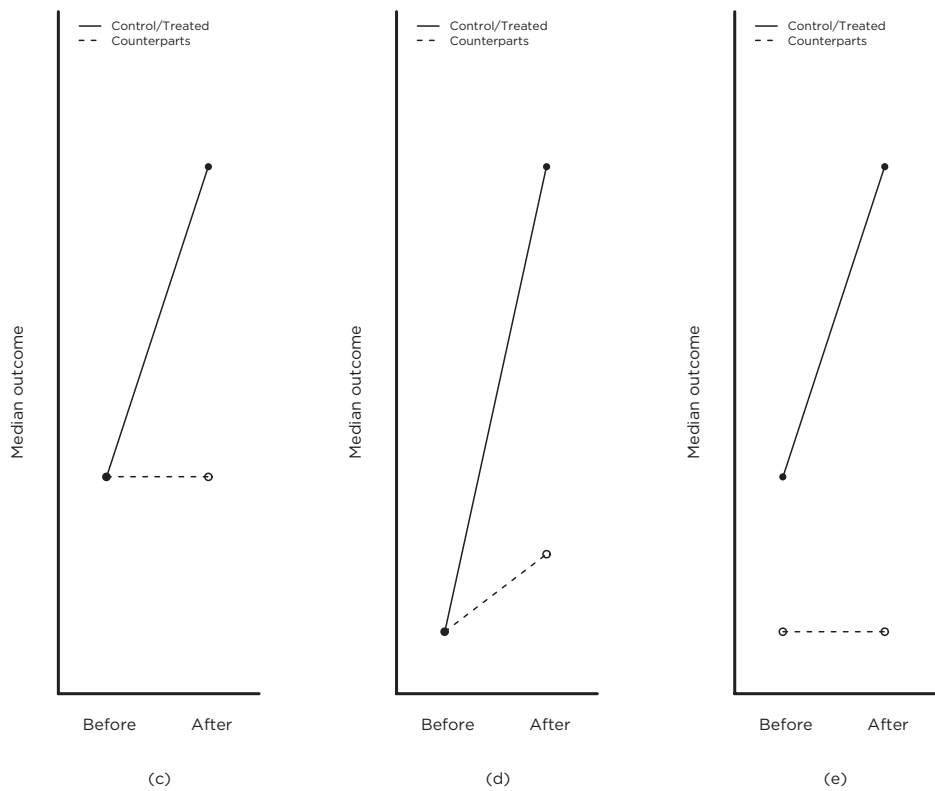
Figure 8.5. Stylized plots with counterparts. There is a treated group in the after period, a control group in the before period, and untreated counterparts. The dots represent the median response in each group; they correspond with the middle line in a boxplot.

The patterns in Figure 8.5 are markedly different from the patterns in Figure 8.4. Taking logarithms in Figure 8.4 changed the qualitative pattern. Specifically, from panel (a) to panel (b); the difference-in-difference changed from a positive number to zero. Taking logarithms in Figure 8.5 would not change the qualitative pattern, although it would squeeze or stretch the distances involved. Panel (a) would look qualitatively similar if logarithms were taken. The same is true of panel (b) and panel (c). Indeed, other increasing transformations, not just the logarithm, would not change the qualitative appearance of Figure 8.5. In this respect, the patterns in Figure 8.5 share a property with causal effects: the patterns and causal effects do not

come and go by taking logarithms. The simple arithmetic of difference-in-differences cannot be fundamental to causal inference if difference-in-differences come and go with transformations but causal effects do not. Panel (a) of Figure 8.4 is less satisfactory than panels (c), (d), and (e) of Figure 8.5 because the pattern in panel (a) might vanish if the data were transformed, say, by taking logarithms, but the patterns in (c), (d), and (e) would persist. To repeat, this is not an argument against examining counterparts; rather, it is an argument against interpreting counterparts in a mechanical way.

Control Outcomes

What Are Control Outcomes?

A control outcome is an outcome that the treatment should not affect. One checks for bias by checking that the treatment is associated with outcomes it might plausibly affect but is not associated with outcomes it should not affect.

More generally, an outcome for which the treatment is known may be used to check for bias by estimating the treatment effect for this outcome, comparing the estimate with what is known. Sometimes only the direction of the effect, not its magnitude, is known. A control outcome is then the special but common case in which the treatment is known to have no effect.

An example was discussed in Chapter 7, in connection with the study by Garren Wintemute and colleagues of the possible effects of restrictions on handgun purchases. It is plausible that such restrictions might affect the rate of crimes for which a handgun is relevant. Restrictions on handgun purchases seem unlikely to reduce the frequency of crimes that normally do not involve a handgun, such as embezzlement or check forgery.

Noel Weiss discussed an example concerning the effectiveness of screening for colon cancer by sigmoidoscopy. Screening was associated with a reduction in mortality from tumors in the region that could be reached by the sigmoidoscope, but with no reduction in mortality from tumors that were beyond its reach.[32] So the treatment effect appears only where an effect is plausible.

Intimate Partner Homicides and Restrictions on Firearms

Elizabeth Richardson Vigdor and James Mercy used control outcomes in a study of the possible effects of laws passed by some U.S. states that restrict access to firearms by individuals who have been convicted of a domestic violence misdemeanor or who are subject to a restraining order.[33] They begin by observing, "Approximately 1 in 3 female homicide victims and 1 in 20 male homicide victims are killed by current or former spouses or boyfriends," and "about 60% of these homicides were committed using a firearm."[34] The study asked whether these laws reduce the rate of intimate partner homicides. The laws vary considerably: some permit or require the police to confiscate firearms at a scene of domestic violence; others prohibit the possession, purchase, or carrying of firearms by individuals convicted of a misdemeanor involving domestic violence. Some of these laws strengthen restrictions already in place under federal law, and others add new restrictions. For example, a federal law restricts access to firearms by individuals subject to a restraining order, but it excludes temporary restraining orders; however, some states have strengthened the federal law by including temporary restraining orders as well. Vigdor and Mercy concluded, "Study results lead us to cautiously conclude that laws restricting access to firearms by abusers under restraining orders reduce [intimate partner homicides]."

A problem with comparisons among states with different laws is the states differ in many ways, including other laws, aspects of law enforcement, and public attitudes about firearms. As a check for unmeasured bias, Vigdor and Mercy included several control outcomes, including rates of "stranger homicides," robberies, and motor vehicle thefts, finding no association of these control outcomes with state laws restricting firearms by individuals subject to a restraining order.[35] The logic here is that a law aimed at individuals subject to restraining orders will have little or no causal effect on these control outcomes; hence, an association between such a law and one of these outcomes is evidence of unmeasured bias distinguishing states that adopt such laws from states that do not.

Checks of this sort can detect some unmeasured biases that lead different states to have different rates of crimes for reasons unrelated to specific laws under study.[36]

Discontinuity Designs

It is common for a government agency to draw a bright red line, saying that people to the left of the line receive treatment and those to the right receive control. Perhaps people who are very far to the left of the line seem well-served by treatment, and people very far to the right of the line seem well-served by control. However, the placement of the red line and the whole idea of a red line seem frustrating and arbitrary: people immediately to the left of the line seem no different from people immediately to the right of the line. Of course, this is just what we wanted: very similar people, some of whom were arbitrarily given treatment and others who were given control. So we compare people who are close to the line but received different treatments by virtue of falling on opposite sides of the line.[37]

The bright red line may be defined by a score on a quantitative assessment. Students who score below a cut-point on an aptitude test are given remedial instruction that is denied to those above the cut-point. In truth, students just above or just below the cut-point are very similar. So the narrow band of students close to the cut-point comprises a basis for evaluating the effects of remedial instruction. In contrast, students far from the cut-point, far above or far below, are quite different as students and are not useful in evaluating remedial instruction.

Additional resources may be made available when a threshold is crossed. In Israel, a rule derived from the medieval scholar Maimonides dictates that an additional teacher must be appointed whenever a class-size exceeds 40 students. Under this rule, enrollment of one additional student may replace one large class by two much smaller classes, and this commonly occurs in Israeli schools. Joshua Angrist and Victor Lavy exploited this fact to study the effects of class size on academic test performance.[38] Their results support the claim that students achieve higher test scores when taught in smaller classes.

The bright red line may appear on a map. People who live on one side of the street may be required to attend a different school than people who live on the opposite side. The two schools may differ in quality. Sandra Black used such circumstances together with house prices to estimate the value parents attach to superior schools.[39]

The bright red line may be defined by the boundaries of voting districts. People who live close to the boundary of a voting district may closely resemble people just across the boundary, but they face different ballots. Luke

Keele, Rocío Titiunik, and José Zubizarreta use this to ask whether the presence of ballot initiatives causes an increase in voter turnout.[40]

Taking Stock

In an observational study, an observed association between treatment received and a measured outcome is ambiguous—perhaps an effect caused by the treatment, perhaps a bias in the way treatments were assigned to individuals. Quasi-experimental devices attempt to reduce ambiguity in a visible association by looking at additional associations, to reduce ambiguity in data by looking at more data. Quasi-experimental reasoning is alert to the plausible alternative explanations of a visible association and therefore alert to the kinds of additional associations that might adjudicate among alternative explanations. Quasi-experimental devices often use readily available data that might otherwise be ignored. More often, these devices distinguish parts or aspects of available data, giving distinct parts distinct roles in a reasoned interpretation of the data. For example, true controls may provide an equitable comparison while counterparts rule out a plausible alternative explanation. The quasi-experimental approach stands opposed to the view that causal inference can successfully proceed mechanically by tossing all the data into a model and turning a crank. Because this mechanical view does not seek to adjudicate among plausible alternative explanations of an association, it leaves such alternatives interpretations intact, including alternatives that might have been invalidated by more thoughtful, less mechanical probing of the data.

Sensitivity to Bias

What Is Sensitivity Analysis?

The General Concept of Sensitivity Analysis

Calculations in mathematics, in engineering, in statistics, and in all of the mathematical sciences depend upon assumptions. What if the assumptions are wrong?

One response, sometimes the correct response, is that the assumptions cannot be wrong, or that the assumptions could only be wrong if something extraordinary occurred. In the ProCESS Trial in Chapter 1, the investigators randomly assigned treatments to individuals using random numbers generated by a computer. In this case, we can safely assume that treatments were indeed assigned at random. This safe assumption leads by logic to the calculations in Chapter 3, to the distribution of our test statistic under the null hypothesis of no treatment effect, to the P-value testing this hypothesis. To claim that treatments were not assigned at random is to claim that the investigators are lying about how they conducted their study. A lie of that magnitude would end a scientist's career. Fraud on that scale is rare, and it is discovered when a particular investigator produces findings that other investigators cannot reproduce.

An engineer is building a bridge to transport trains across a river. The engineer makes various assumptions and from these assumptions various calculations, and on the basis of these calculations the engineer determines how much steel is needed, how deep the pilings must be, and so on. What if the assumptions are wrong? Henry Petroski wrote that, in engineering, "success is foreseeing failure," a useful thought in many contexts.[1] In building the bridge, the engineer would create a margin of safety in the assumptions, allow for a reasonable but substantial failure of plausible assumptions, by building the bridge to withstand stresses much greater than any that are likely to occur. Such a margin of safety is reasonable in building bridges. A margin of safety is justified by weighing the consequences of various mistakes, preferring to waste a little steel rather than risk many deaths.

Margins of safety are not useful in contexts in which there is no way to err on the side of safety, contexts in which errors in either direction could do grave harm. If we are ignorant about whether a new drug does more good than harm, it may be unclear how to err on the side of safety.

A sensitivity analysis is an analysis directed at a calculation that depends upon assumptions that may be false. A sensitivity analysis asks, How would the results of the calculation, or the conclusion, change if the assumptions were changed by a limited amount? Would the conclusion barely change? If so, the conclusion is insensitive to a violation of the assumptions of that limited magnitude. In contrast, if the conclusions would change substantially when the assumptions are changed by a limited amount, then the conclusions are sensitive to a violation of the assumptions of that limited amount. As an expedient, we often ask, How large would the violation of the assumptions have to be to materially alter the conclusion? What would it take for the bridge to collapse? Sensitivity analyses are done because they are often enlightening. A sensitivity analysis may provide grounds for caution that are not rooted in timidity, or grounds for boldness that are not rooted in arrogance. John Dewey wrote, "It is the *situation* that has these traits. We are doubtful because the situation is inherently doubtful. Personal states of doubt that are not evoked by and are not relative to some existential situation are pathological; when they are extreme they constitute a mania of doubting."[2]

In Chapter 5, we saw that if treatment assignment were ignorable with observed covariates x_i, then relatively straightforward adjustments for x_i, such as matching treated and control subjects with the same value of x_i, would suffice to produce correct causal inferences. We also saw that in observational

studies we rarely have compelling reasons to believe the assumption that treatment assignment is ignorable. In Chapter 6, we saw that violations of ignorable treatment assignment resemble descending from a summit along a gradual slope, not falling off a cliff. In seeking natural experiments, we seek to be as close as possible to the summit. A sensitivity analysis in an observational study asks how far we would have to descend from the summit before materially changing the conclusions. It asks about the magnitude of the violation of ignorable treatment assignment that would have to be present to materially alter the conclusions of the study.

What Is Not a Sensitivity Analysis?

The term "sensitivity analysis" is sometimes misused. Some activities that are not sensitivity analyses are mistakenly described as sensitivity analyses. Sensitivity analysis is a venerable term from the mathematical sciences, and it is targeted at the consequences of violations of assumptions that are integral to a particular mathematical calculation. If you stop and think about how the mathematical sciences work, you realize that violations of assumptions are a concern whenever mathematical reasoning is applied to practical problems. What if the assumptions are wrong?

Sometimes an investigator with limited training in statistics is impressed by the many statistical analyses that modern software can easily perform. Unsure which analysis is appropriate under what circumstance, this investigator performs many analyses, reports a few that support the investigator's preferred conclusion, and calls the activity a sensitivity analysis. When this occurs, it is a misuse of the term sensitivity analysis. This kind of activity is strongly discouraged in randomized clinical trials, where a protocol written in advance of the trial specifies one primary statistical analysis. In a sharp critique of observational studies that shop among analyses for preferred conclusions, Donald Rubin advocated this position:

> Observational studies can and should be designed to approximate randomized experiments as closely as possible. In particular, observational studies should be designed using only background information to create subgroups of similar treated and control units, where 'similar' here refers to their distributions of background variables [i.e., covariates]. Of great importance, this

activity should be conducted without any access to any outcome data, thereby assuring the objectivity of the design . . . Of course, objectivity is not the same as finding truth, but I believe that it is generally a necessary ingredient if we are to find truth.[3]

The argument for one fixed primary analysis specified before examining outcomes is the same in randomized experiments and observational studies: it precludes the possibility that the investigator shops among analyses for a preferred conclusion.

A sensitivity analysis in an observational study varies a key assumption in one primary statistical analysis, thereby displaying the degree to which violations of that assumption would destabilize the study's conclusions. A sensitivity analysis does not perform many disconnected analyses, and it certainly does not encourage the dishonest practice of performing many disconnected analyses and reporting only a favored few.[4]

The First Sensitivity Analysis in an Observational Study

The first sensitivity analysis in an observational study occurred in an article by Jerome Cornfield and colleagues.[5] The article appeared in the *Journal of the National Cancer Institute* in 1959, and it appraised evidence suggesting that cigarette smoking causes cancer of the lung. Because of its influence, the article was republished a half century later in 2009 in the *International Journal of Epidemiology* with commentary by David Cox, Jan Vandenbroucke, Marcel Zwahlen, and Joel Greenhouse.[6]

Cornfield and colleagues wrote,

If an agent, A, with no causal effect upon the risk of a disease, nevertheless, because of a positive correlation with some other causal agent, B, shows an apparent risk, r, for those exposed to A, relative to those not so exposed, then the prevalence of B, among those exposed to A, relative to the prevalence among those not so exposed, must be greater than r. Thus, if cigarette smokers have 9 times the risk of nonsmokers for developing lung cancer, and this is not because cigarette smoke is a causal agent, but only because cigarette smokers produce hormone X, then the proportion of hormone-X-producers among cigarette smokers must be at least 9 times greater than that of nonsmokers.[7]

This statement is an important conceptual advance. We know that "association does not imply causation," that associations between treatment received and outcome exhibited may be due to biases in the way treatments are assigned. The quoted statement by Cornfield and colleagues replaces this true statement by another true statement that is more useful—a quantitative statement that makes use of the observed data—namely, to explain an observed association of a particular magnitude, the bias in treatment assignment must exceed another particular magnitude.

Joel Greenhouse summarizes the quoted statement in this way: "No longer could one refute an observed causal association by simply asserting that some new factor (such as a genetic factor) might be the true cause. Now one had to argue that the relative prevalence of this potentially confounding factor was greater than the observed relative risk of the putative causal agent."[8] In other words, this first sensitivity analysis transferred some of the onus from the investigator to the critic. The critic could no longer say, "Anything can explain everything." Not just any criticism would do: the critic's imagined unobserved covariate u_i had to meet certain standards to explain away the observed association. Wittgenstein asked, "Doesn't one need grounds for doubt?"[9] Aided by this sensitivity analysis, the grounds for doubt might themselves be judged doubtful, inadequate, unreasonable, outlandish, outrageous, or not, depending upon the observed association and the scientific context.

Irwin Bross put the matter this way:

> A critic who objects to a bias in the design or a failure to control some established factor is, in fact, raising a counterhypothesis . . . What is the responsibility of a critic with respect to his counterhypothesis? . . . The critic has the responsibility for showing that his counterhypothesis is tenable. In so doing, he operates under the same ground rules as a proponent.[10]

Limitations of the First Method of Sensitivity Analysis

Though an important conceptual advance, the specific method proposed by Cornfield and colleagues is of limited applicability, giving rise to a large literature designed to address its limitations.[11] The original method was restricted to binary outcomes without observed covariates, and because it did

not take account of sampling variability it could be misleading except in very large studies. This chapter briefly discusses one approach to sensitivity analysis that is similar in spirit to the method introduced by Cornfield and colleagues.[12]

Quantifying Departures from Randomized Treatment Assignment

Where Do Things Stand So Far?

Chapter 3 discussed testing Fisher's null hypothesis, H_0, of no treatment effect in a completely randomized experiment. Fisher's hypothesis asserted that each person i would exhibit the same response if given treatment with $Z_i = 1$ or if given the control with $Z_i = 0$; that is, it asserted the two potential outcomes for person i, namely r_{Ti} under treatment and r_{Ci} under control, were equal. For instance, in the ProCESS Trial, patient $i = 17$, Harry, was given aggressive treatment, $Z_{17} = 1$, and was observed to survive, $R_{17} = r_{T17} = 0$, and Fisher's hypothesis says that Harry would also have survived had he instead been given the less aggressive treatment, $r_{C17} = 0$. In Table 3.1, there were $T = 92$ deaths in the aggressive treatment group, and Chapter 3 asked whether seeing these 92 deaths provided much evidence against the hypothesis of no effect, concluding that it did not. Specifically, Chapter 3 used two elements—Fisher's hypothesis and random assignment of treatments—to deduce the behavior or distribution of the statistic, T, when the null hypothesis is true. Because seeing 92 deaths in the aggressive group would have been unremarkable were the null hypothesis true, we concluded that the data provide little or no evidence suggesting that the null hypothesis is false. In the ProCESS Trial, there was nothing speculative about this calculation because we knew that treatments were randomly assigned by the ProCESS Investigators. In brief, the key assumption—that treatment assignment was randomized—was ensured by the use of random assignment in the design of this clinical trial.

In Chapters 5 through 8, similar calculations were done in four observational studies: in the study of delirium in Table 5.5, in the study of posttraumatic stress in Table 6.3, in the study of lead in children's blood in Figure 7.2, and in the study of the duration of injury compensation in

Figure 8.2. In each of these cases, the calculated *P*-value would have been correct were it true that treatment assignment is ignorable given observed covariates, x_i, but we had little basis for making this key assumption. Would small violations of this assumption radically change our interpretation of these studies? Or would it take very large violations of ignorable treatment assignment to render plausible the null hypothesis of no treatment effect?

Each of these four observational studies had a simple form: it consisted of matched pairs of one treated individual and one control who had the same or similar values of the observed covariate x_i. The study of lead in children's blood and the study of the duration of injury compensation had additional information, either doses of exposure or counterparts, but let us start by considering the common element in the four studies, the treated–control matched pairs.

Recall that ignorable treatment assignment means that if two people, say, person *i* and person *j*, look the same, if they have the same value of the observed covariate, $x_i = x_j$, then they have the same probability of treatment, $\pi_i = \pi_j$. That may or may not be true. In the discussion of matched pairs in Chapter 5, we formed pair *p* by matching two unrelated people with the same observed covariate, say, person *i* and person *j* with $x_i = x_j$, one of whom was treated and the other control, so that $Z_i + Z_j = 1$. We saw in Chapter 5 that if treatment assignment is ignorable, then this matching has fixed the problem: in the resulting matched pair, person *i* and person *j* are equally likely to be the one treated person in pair *p*. Let us state this same thought a little more precisely. Let us write θ_p for the probability that, in the resulting pair *p*, it is the first person *i* who is treated, with $Z_i = 1$ and $Z_j = 0$; so $1 - \theta_p$ is the probability that it is the second person *j* who is treated, with $Z_i = 0$ and $Z_j = 1$.[13] Matching ensured $x_i = x_j$ in pair *p*, and if treatment assignment is ignorable this implies $\pi_i = \pi_j$; so requiring exactly one treated person in pair *p*—that is, requiring $Z_i + Z_j = 1$—means $\theta_p = 1/2$. In Chapter 5 this was demonstrated by an argument from symmetry; in Chapter 5, note 19, it was demonstrated by an argument from conditional probability.

Because this is an important step, let us do it one more time. For some reason, we did not record gender in *x*. Suppose that we have paired Harry, $i = 1$, and Sally, $i = 2$, because they have the same value of the observed co-variates, $x_1 = x_2$, and either Harry or Sally was treated but not both, $Z_1 + Z_2 = 1$. In this first pair, $p = 1$, the chance that Harry is treated and Sally is not—the chance that $Z_1 = 1$ and $Z_2 = 0$ is θ_1—and the chance that Sally is treated and

Harry is not—the chance that $Z_1 = 0$ and $Z_2 = 1$—is $1 - \theta_1$; it has to be one or the other because of the way we constructed the pair. If treatment assignment is ignorable—a big if—then our problems are solved: $\theta_1 = 1/2$, and whether Harry or Sally received treatment in pair $p = 1$ is just a fair coin flip.

The next section will use θ_p to quantify violations of ignorable treatment assignment.

Measuring Departures from Ignorable Treatment Assignment in Matched Pairs

In treatment–control pairs matched for x_i, the chance that the first person in pair p is treated is $\theta_p = 1/2$ under the assumption that treatment assignment is ignorable. What if that assumption is wrong?

At several points, we have spoken of departures from ignorable assignment as descending from a summit along a gradual slope, rather than as falling off a cliff. Falling off a cliff would mean this: if I cannot assume the probability θ_p is $\theta_p = 1/2$, then I must assume θ_p can take any value between 0 and 1. Descending a gradual slope simply means taking an interest in values of θ_p that depart from $1/2$ by a limited amount. To perform a sensitivity analysis, we do not need to know by how much θ_p departs from $1/2$—we cannot know this, so we had better avoid a method that presumes we know what we cannot know. Rather, we can suppose θ_p could depart from $1/2$ by a modest amount. Would the conclusions change? Suppose θ_p could depart from $1/2$ by a larger amount. Now, would the conclusions change? We might speed up the process. It is often clearer to us what it means for the conclusions to change—what it means for the bridge to collapse—than it is what values of θ_p are reasonable. So we could speed things up by working backward, and ask, By how much would θ_p have to depart from $1/2$ in order to change the conclusions? Perhaps we have rejected, at the conventional 0.05 level, Fisher's hypothesis of no treatment effect in a calculation that assumes treatment assignment is ignorable; that is, we obtained a P-value of 0.05 or less. In fact, we did this in each of the four paired examples, delirium, lead, post-traumatic stress, and injury compensation. By how much would θ_p have to depart from $1/2$ to obtain a P-value above 0.05 so that we can no longer reject the hypothesis of no effect? Or perhaps we have an estimate of the treatment effect, and it is positive, say, beneficial. By how much would θ_p

Table 9.1.   Understanding the sensitivity parameter Γ

Γ	Range of possible values of θ_p		Λ	Δ
1	0.50	0.50	1	1
1.05	0.49	0.51	1.37	1.37
1.1	0.48	0.52	1.40	1.80
1.25	0.44	0.56	2	2
1.5	0.40	0.60	2	4
2	0.33	0.67	3	5
2.5	0.29	0.71	4	6
3	0.25	0.75	5	7
3.5	0.22	0.78	6	8
4	0.20	0.80	7	9
4.5	0.18	0.82	8	10
5	0.17	0.83	9	11
6	0.14	0.86	11	13
7	0.12	0.88	13	15
8	0.11	0.89	15	17
9	0.10	0.90	17	19
10	0.09	0.91	19	21

have to depart from 1/2 to obtain an estimate of an effect that is negative, say, harmful? We can ask the same question of confidence intervals, of equivalence tests, and of multiple testing procedures such as the three-sided test. Once we have the correct tools, we can answer each of these questions in a single sentence.

Table 9.1 is a single yardstick for measuring departures from ignorable treatment assignment in matched pairs. Each row is a specific distance from ignorable assignment. There are several numbers in each row, but these should be understood as analogous to a yardstick with inches on the left side and centimeters on the right. Some people like inches, and others like centimeters, but the distances are in perfect correspondence. You are more centimeters tall than you are inches tall, but your height is the same. There are reasons to like each of the marks on the yardstick in Table 9.1, and each is easy to understand, so consider them one at a time.

The first row of Table 9.1 is ignorable treatment assignment, the summit, and in that row the range of possible values of θ_p is from $0.50 = 1/2$ to $0.50 = 1/2$, so $\theta_p = 1/2$. In the second row, θ_p is very close to 1/2, perhaps as

low as 0.49, perhaps as high as 0.51. This is a small but not completely trivial departure from ignorable assignment. In the third row, θ_p can be as low as 0.48 or as high as 0.52. If you were flipping coins that came up heads with probability $\theta_p = 0.48$ and tails with probability $1 - \theta_p = 0.52$, then you might discover this, but only after flipping the coin many times. In the bottom row of Table 9.1, the probability θ_p can be as low as 0.09 or as high as 0.91, not quite anything, 0 to 1, but almost anything.

As ticks on a yardstick, there is nothing terribly wrong with the interval for θ_p, but the number Γ on the left side is a bit better. For example, $\Gamma = 2$ is the same magnitude of departure from ignorable treatment assignment as the interval from $1/3 \approx 0.33$ to $2/3 \approx 0.67$ for θ_p. Let us first understand what Γ is, and once that is clear, turn to why Γ is a bit better. Notice first that $1/3 = 1/(1+2)$ and $2/3 = 2/(1+2)$. In general, the interval for θ_p may be expressed in terms of Γ as

$$\frac{1}{1+\Gamma} \leq \theta_p \leq \frac{\Gamma}{1+\Gamma}.$$

As a second example, $\Gamma = 3$ is the same as the interval from $0.25 = 1/4 = 1/(1+3)$ to $0.75 = 3/(1+3)$.

Not everyone talks about probabilities. Casinos and bookmakers have a practical interest in probabilities, but they describe them in terms of odds. A probability of 1/2 is described as "even money" or 1-to-1 odds. A coin flip has 1-to-1 odds or $1/1$, meaning that the two outcomes are equally likely. A probability of $2/3$ has 2-to-1 odds or $2/1$. A biased coin with probability $2/3$ of a head is twice as likely to come up heads as tails, 2-to-1. Roll a die, and the odds against getting a 1 are 5-to-1 or $5/1$, and this corresponds with a $5/6$ probability of not getting a 1 and a $1/6$ probability of getting a 1. If you know the odds, you can recognize a fair bet. When rolling a die, a bet of $5 that you will not get a one against $1 that you will is a fair bet; neither side has an advantage on average. Casinos and bookmakers are interested in fair bets because they never gamble; rather, they offer many people unrelated subfair bets and are almost certain to win in aggregate. To offer people a subfair bet, you have to be able to recognize a fair bet and offer less; that is, you have to know the odds. For a probability θ_p the odds are $\theta_p/(1 - \theta_p)$. A probability of $\theta_p = 2/3$ is an odds of $(2/3)/(1 - 2/3) = 2/1$ or 2-to-1.

In Table 9.1, Γ is a bound on the odds. Remember, Harry and Sally were paired to form pair $p = 1$ because they look the same in terms of x_i and exactly one of them received treatment. If $\Gamma = 2$, then Harry might be twice as likely as Sally to receive treatment, or Sally might be twice as likely as Harry to receive treatment; that is, the odds, $\theta_1 / (1 - \theta_1)$, are at most $\Gamma = 2 / 1$ and at least $1 / \Gamma = 1/2$. Now the odds, $\theta_1 / (1 - \theta_1)$, are $2 / 1$ or 2-to-1 if $\theta_1 = 2 / 3$, and the odds are $1/2$ or 1-to-2 if $\theta_1 = 1 / 3$, as in Table 9.1. In general, the relationship between the possible odds $\theta_p / (1 - \theta_p)$ and Γ is given by the inequality

$$\frac{1}{\Gamma} \leq \frac{\theta_p}{1 - \theta_p} \leq \Gamma.$$

With a little algebra, you can rearrange this inequality for the odds $\theta_p / (1 - \theta_p)$ back into the previous inequality for the probability θ_p; that is, the two inequalities say the same thing in different ways.

As alternative ticks on the same yardstick, why is Γ better than an interval for θ_p? There are two reasons, the second being more important. The first reason is that the interval for θ_p is expressed in two numbers, while Γ is expressed in one number, yet they convey the same information. Just as it is more natural to say, "I put my shoes on," rather than to say, "I put my left shoe on and my right shoe on," so too it is more natural to speak in terms of Γ rather than the interval for θ_p. It is a concise way of saying the same thing.

The second and more important reason for preferring Γ to an interval for θ_p is that θ_p is tied to matched pairs, but Γ can be used to describe many situations. If we had seen an unmatched 2×2 table like Table 3.1 in an observational study, we could not measure departures from ignorable treatment by θ_p because θ_p is only defined for matched pairs, but we could still use Γ. If we had a stratified table like Table 5.1, we could not measure departures from ignorable treatment by θ_p, but we could still use Γ. If we had a study like the crying babies example in Chapter 5, with one treated subject and several controls in each of many matched sets, then we could not measure departures from ignorable treatment by θ_p, but we could still use Γ. In Chapter 8, in the example involving injury duration, we did two tests, one for treated-control pairs, the other comparing the treated-minus-control

differences to the counterpart differences. We can use the interval for θ_p to think about the first comparison, the treated–control pairs, but we would need Γ to think about the unmatched comparison with counterparts. In this book, I will do sensitivity analyses only for matched pairs, so you can use the interval for θ_p if you prefer, but once you step out of this book into the larger world, Γ is handy.[14]

The final two columns of Table 9.1 are described in the following starred, optional section.

* Amplification of Sensitivity Analysis

The sensitivity analysis of Cornfield and colleagues and the analysis defined in terms of either θ_p or Γ both speak of departures from ignorable treatment assignment exclusively in terms of the probability of treatment, π_i. It is sometimes convenient to speak of the outcome, (r_{Ti}, r_{Ci}), when discussing departures from ignorable treatment assignment. Fortunately, this alternative way of speaking is simply a new set of marks on the same yardstick, a new unit of measure besides inches and centimeters, not a new quantity being measured.[15] The new marks are the final two columns in Table 9.1.

Suppose that Fisher's null hypothesis of no treatment effect is true, so $R_i = r_{Ti} = r_{Ci}$ for each person i. Then any association between the outcome $R_i = r_{Ci}$ and the treatment Z_i is spurious, a bias in treatment assignment, not an effect caused by the treatment. To create a spurious association after matching for the observed covariates x_i, the unobserved covariate u_i must persist in being related to both r_{Ci} and Z_i. Both the sensitivity analysis of Cornfield and colleagues and the analysis in terms of Γ permit a very strong relationship between r_{Ci} and u_i; then they limit the relationship between Z_i and u_i. So both analyses measure departures from ignorable treatment assignment by a single number, and this is often expedient, as in the first column of Table 9.1.

Sometimes, however, the conversation is about a specific unobserved covariate u_i about which people have strong opinions that they want to express in the sensitivity analysis. For example, suppose that a study of the health effects of heavy smoking, Z_i, did not record the level of alcohol consumption, and the entire debate is about the unobserved covariate u_i recording mean daily alcohol consumption. On average, in the United States smokers

tend to drink more than nonsmokers, but of course many people drink without smoking or smoke without drinking, so there is a relationship between Z_i and u_i, but it is far from a perfect relationship. Suppose we were interested in two different health outcomes, lung cancer and cirrhosis of the liver. Now there is good reason to think that alcohol consumption, u_i, is strongly related to cirrhosis of the liver, and not much reason to think it is nearly as strongly related to lung cancer. So we have much more reason to worry about this specific u_i when studying the outcome of cirrhosis than when studying the outcome of lung cancer. At first appearances, it seems that both the sensitivity analysis of Cornfield and colleagues and the analysis in terms of Γ lack the ability to speak to this situation, because they both speak about the relationship between Z_i and u_i and do not mention a specific outcome.

As it turns out, this initial appearance is not correct. For the method of Cornfield and colleagues, this was demonstrated by Joseph Gastwirth, not by introducing a new method but by reinterpreting the components of the existing method.[16] The same thing can be done with Γ, and this occurs in the last two columns of Table 9.1.[17] Again, the last two columns are a reinterpretation of Γ; they are new marks on an old yardstick, not a new quantity being measured.

The last two columns of Table 9.1 refer to a spurious association, that is, an association between treatment received, Z_i, and outcome observed, R_i, in the absence of a treatment effect, $R_i = r_{Ti} = r_{Ci}$, created by a single unobserved covariate, u_i, by virtue of relationships between Z_i and u_i and between r_{Ci} and u_i. Person i and person j have been paired to form pair p by virtue of having the same observed covariates, $x_i = x_j$, and one treated subject, $Z_i + Z_j = 1$; this simplifies the discussion by getting the observed covariates out of the picture. Keeping in mind that there is no treatment effect, write $Y_{Cp} = R_i - R_j = r_{Ci} - r_{Cj}$ for the difference in outcomes in this pair p, and $V_p = Z_i - Z_j = 2Z_i - 1$ for the difference in treatments, $V_p = \pm 1$. Again, there is a spurious association between Y_{Cp} and V_p because both depend upon (u_i, u_j); that is, the pair difference in treatments, V_p, predicts the pair difference in outcomes, Y_{Cp}, in the absence of a treatment effect because both V_p and Y_{Cp} depend on the unobserved covariate. The parameter $\Lambda \geq 1$ defines the strength of the relationship between V_p and (u_i, u_j): the odds of $V_p = 1$ rather than $V_p = -1$ are at most Λ and at least $1 / \Lambda$.[18] The parameter $\Delta \geq 1$ defines the strength of the relationship between Y_{Cp} and (u_i, u_j): the odds of $Y_{Cp} > 0$ rather than $Y_{Cp} < 0$ are at most Δ and at least $1 / \Delta$.

In the situation just described, the impact of (Λ,Δ) is the same as the impact of Γ whenever $\Gamma = (\Lambda\Delta + 1)/(\Lambda + \Delta)$.[19] For example, in Table 9.1, $\Gamma = 1.5$ is the same as $(\Lambda,\Delta) = (2,4)$ because $1.5 = (2 \times 4 + 1)/(2 + 4)$. In other words, an unobserved covariate u that could double the odds of treatment in a matched pair and could increase the odds of a positive pair difference in responses by a factor of 4 is the same as $\Gamma = 1.5$. This is, however, only an illustration: any (Λ,Δ) that solves $1.5 = (\Lambda\Delta + 1)/(\Lambda + \Delta)$ is also equivalent to $\Gamma = 1.5$. For instance, $\Gamma = 1.5$ is the same as $(\Lambda,\Delta) = (2,4)$ and $(\Lambda,\Delta) = (4,2)$ and $(\Lambda,\Delta) = (2.5, 2.75)$.

In the previous paragraph, each single value of Γ is replaced by a curve consisting of all values of (Λ,Δ) that solve $\Gamma = (\Lambda\Delta + 1)/(\Lambda + \Delta)$. This replacement of each Γ by a different curve is called an "amplification." Table 9.1 picked a representative pair (Λ,Δ) from a curve with infinitely many choices of (Λ,Δ). The particular values in Table 9.1 were selected for aesthetic reasons—a preference for integer values when possible, and values of Λ and Δ not too dissimilar in size. A particular scientific context might dictate other preferences.

Expressing Γ equivalently as (Λ,Δ) by solving $\Gamma = (\Lambda\Delta + 1)/(\Lambda + \Delta)$ is convenient. An investigator can perform and report a single sensitivity analysis in terms of Γ, yet without further analysis or computation have available the various corresponding interpretations in terms of (Λ,Δ). For example, in the case discussed earlier, if u were alcohol consumption and Z were heavy smoking, then we would be interested in larger values of Δ if the outcome were cirrhosis of the liver than if the outcome were lung cancer. Alternatively, an unobserved covariate may be strongly related to treatment assignment, but its relationship with the outcome is thought to be weak, so Λ could be large with Δ much smaller; however, one does not need a new sensitivity analysis to consider this case. Finally, the equivalence $\Gamma = (\Lambda\Delta + 1)/(\Lambda + \Delta)$ is another way of understanding the magnitude of Γ; so $\Gamma = 1.25$ corresponds to an unobserved covariate that doubles the odds of treatment, $\Lambda = 2$, and doubles the odds of a positive pair difference in responses, $\Delta = 2$.

A Sensitivity Analysis

Lead in Children's Blood

Chapter 7 discussed a study by Morton and colleagues of lead in the blood of children whose parents were exposed to lead at work.[20] Did the parent's

exposure to lead cause an increase in the child's level of lead? The results of
that study were depicted in Figures 7.2 and 7.3. Here, we will focus on the
treated-minus-control pair differences in lead levels, the comparison in panel
(a) of Figure 7.2 or in panel (a) of Figure 7.3. A sturdy, old, conventional
test, Wilcoxon's signed rank test, looks at the pair differences, finds that most
differences are positive, and concludes that seeing so many large positive pair
differences would have been very improbable in a paired randomized experi-
ment in which the treatment has no effect.[21] Specifically, the one-sided
P-value testing Fisher's hypothesis of no treatment effect would be 5.5×10^{-6}
had this been a randomized experiment, and the two-sided P-value would
be twice that value.

The calculation in the previous paragraph leaves us in something of a
quandary. It says that Figures 7.2, panel (a) and 7.3, panel (a) would be com-
pelling evidence that parental exposures caused an increase in their
children's lead levels had the data come from a randomized experiment.
However, we know that the data did not come from a randomized experi-
ment. To assume that the data came from a randomized experiment when
that is not the case is to assume something that is not true. If we assume
something that is not true, we may deduce from that false assumption many
other things that are not true. So we ask, How sensitive is this conclusion to
violations of the assumption that we know is untrue? If we permitted small
violations of the assumption, as measured in Table 9.1, would the conclu-
sion be completely different? Or would it take very large violations of this
assumption to materially alter the conclusion?

Sensitivity Analysis for a P-Value

The P-values in the previous paragraph assumed ignorable treatment assign-
ment, essentially random assignment with pairs, so that $\theta_p = 1/2$ in every
pair p. A straightforward calculation shows that every possible pattern of
treatment assignment probabilities θ_p with $\Gamma = 4.4$ would produce a one-sided
P-value of 0.049 or less, and every pattern with $\Gamma = 3.5$ would produce a two-
sided P-value of 0.047 or less.[22] In other words, the hypothesis of no treatment
effect would be rejected with a one-sided P-value below the conventional
level of 0.05, providing the assignment probabilities θ_p in the pairs were in
the range

$$\frac{1}{1+\Gamma} = \frac{1}{1+4.4} = 0.185 \leq \theta_p \leq 0.815 = \frac{4.4}{1+4.4} = \frac{\Gamma}{1+\Gamma}$$

This is a large departure from a randomized experiment—instead of every θ_p being 1/2, the θ_p can be any probabilities in the interval [0.185, 0.815], yet the main conclusion would stand.

Using the last two columns of Table 9.1, or the amplification discussed in the starred section, the value $\Gamma = 4.4$ is seen to correspond with an unobserved covariate u_i that produces an eightfold increase in the odds of treatment and a 9.5-fold increase in the odds of a positive pair difference in lead levels.[23] So to claim that the hypothesis of no effect should not be rejected at the 0.05 level, one must postulate quite an impressive unobserved covariate, one that strongly predicts which of two neighbors will work in the battery plant, and also which of two children will have higher levels of lead in their blood. This is, again, as in Chapter 7, the answer to Wittgenstein's question: "What would a mistake here be like?"[24] A mistake in rejecting no effect at the 0.05 level would entail the failure to match for a covariate u_i of considerable importance to both the treatment assignment and the outcome.

Which Test Statistic Is Best?

Wilcoxon's test is a sturdy, familiar, old test. Is it a good choice of test for sensitivity analysis? It turns out that there are better choices. Wilcoxon's statistic has some attractive properties when used in randomized experiments with small treatment effects, but small treatment effects tend to be sensitive to small biases in observational studies. Focusing on small effects is a mistake when small biases can easily explain small effects. In large observational studies, Wilcoxon's statistic tends to exaggerate the degree of sensitivity to unmeasured biases. Because it cares so much about small effects, Wilcoxon's statistic often fails to adequately notice evidence that some effects are not small and are not easily dismissed as produced by small biases. Better test statistics tend to say that conclusions are insensitive to larger biases when, in fact, there is a treatment effect without bias. These better test statistics ignore treated-minus-control pair differences Y_p that are near zero, whether positive or negative. When applied to the lead data, one such statistic yields

a one-sided P-value of 0.05 for $\Gamma < 6$, rather than $\Gamma = 4.4$, a substantial differ-ence.[25] A bias of $\Gamma = 6$ is very large: it would explain away the effect of heavy smoking on lung cancer in one of the major studies of the subject.[26] Again, the change from $\Gamma = 4.4$ to $\Gamma = 6$ is produced by simply selecting a better test statistic.

Sensitivity Analysis for Estimates and Confidence Intervals

The discussion has focused on the sensitivity analysis for a P-value, but similar methods apply to estimates and confidence intervals. For instance, a standard estimate of the increase in blood lead levels due to treatment is 15.41 µg/dl—approximately the center line in Figure 7.6, panel (a)—with a 95% confidence interval [10.14, 21.60]. If we allow for a bias of magnitude $\Gamma = 2$, the one estimate 15.41 is replaced by an interval of estimates, [11.32, 20.18], and the 95% confidence interval becomes longer: [5.12, 27.86]. In words, even allowing for the small sample size and a substantial bias in treatment assign-ment, $\Gamma = 2$, there would still be evidence of a substantial treatment effect.[27]

Analyses of Other Studies

What Do We Learn by Comparing Sensitivity to Bias in Different Studies?

Several things happen when we compare the results of sensitivity analyses in several or many studies. First, we observe that studies differ markedly in their sensitivity to unmeasured biases. Some studies would be destabilized by small biases, whereas others resist very large biases. As one consequence, the sensitivity analysis tells us something about the studies that we did not know before doing the analysis. When you pretend that a large observational study is a large randomized experiment, you are often misled to believe that its conclusions are extremely solid, yet these conclusions may be sensitive to tiny biases from nonrandom treatment assignment.

Second, sensitivity to unmeasured bias does not closely track other quan-tities we often calculate, such as P-values testing no effect in randomized

experiments, or estimates of the size of the treatment effect; so you cannot look at these familiar quantities and know from them whether the conclusions are insensitive to small biases. If you want to know the results of a sensitivity analysis, then you have to perform the sensitivity analysis. A *P*-value computed assuming that there are no unmeasured biases may be small and significant because, in fact, there are unmeasured biases. A confidence interval for a treatment effect computed assuming there are no unmeasured biases may be short and far from no effect because, in fact, there are unmeasured biases. You cannot perform analyses assuming that there are no unmeasured biases, yet learn from these analyses what would happen if unmeasured biases were present. You cannot gain insight into a problem by starting with the assumption that the problem is absent.[28] The story is told of an investigator who escaped from a deep pit by assuming a ladder. If you think that strategy might work, may I suggest that you find enlightenment at the bottom of a deep pit.

Third, we begin to ask, What, exactly, makes some studies insensitive to large biases and other studies sensitive to small ones? We find that certain worldly situations, certain study designs, and certain methods of analysis reliably produce insensitive conclusions, whereas other situations, designs, and methods reliably produce sensitive conclusions. We are therefore attracted to find situations, choose study designs, and employ methods that produce insensitive conclusions. Knowing what produces success, we plan investigations that are likely to succeed. To make more than superficial progress at this third task, we need to ask what will happen in situations we understand because we created these situations to glean understanding. So this third aspect leads to the topic of Chapter 10, design sensitivity, the sensitivity to unmeasured bias in a large sample from a well-defined sampling situation.

Delirium in the Hospital

Chapter 5 considered the study by Sharon Inouye and colleagues of an intervention intended to reduce the frequency of delirium among elderly patients when hospitalized.[29] The data described 852 treated-versus-control matched pairs and appeared in Table 5.5, where delirium was less common in the treated group than among matched controls. When the

data were analyzed as if they had come from a randomized trial, the hypothesis of no treatment effect yielded a one-sided P-value of 0.0123 and a two-sided P-value of $2 \times 0.0123 = 0.0246$. How sensitive are these findings to the false assumption that the data came from a paired randomized experiment?

At $\Gamma = 1.14$, the largest possible one-sided P-value is 0.051. A bias of $\Gamma = 1.14$ is just about equal to a bias produced by an unobserved covariate, u_i, which doubles the odds of delirium and increases the odds of treatment by 50%. So this finding is insensitive to small biases but is much more sensitive to bias than the study of lead discussed in the previous section.[30]

The study of delirium is sensitive to much smaller biases than the study of lead in children. Does this mean that the study of delirium reached the wrong conclusion? It does not. Sensitivity to bias of a particular magnitude, Γ, does not mean that bias of this magnitude is or is not present.[31] The two sensitivity analyses simply say that smaller biases could explain the results of the delirium study, not that such biases do explain the results. Remember, a sensitivity analysis puts some of the burden of proof back on a study's critics—the alternative explanations they offer in terms of u_i must be plausible despite the magnitude of bias needed to alter the study's findings. The sensitivity analysis transferred to the critic a much larger burden of proof in the study of lead than in the study of delirium. Conversely, there have been a few studies that were insensitive to large biases but were mistaken in their conclusions, because enormous biases were present.[32] Sensitivity to unmeasured bias is a consideration in causal inference; it is not the only consideration.

Injury, Compensation, and Incentives

Chapter 8 discussed the study by Bruce Meyer, Kip Viscusi, and David Durbin of the effect of an increase in injury compensation on the duration of time out of work.[33] Consider the matched pair differences in durations depicted in Figure 8.3 for workers with higher wages, those affected by the change in the benefits ceiling. Had these differences been seen in a paired randomized experiment, then the one-sided P-value testing the null hypothesis of no effect using Wilcoxon's statistic would have been 2.55×10^{-7}, and the two-sided P-value would have been twice that. How sensitive is this

finding to violations of the false assumption that the data came from a paired randomized experiment?

At $\Gamma = 1.27$, the largest one-sided P-value is 0.054. In other words, the study falls between the study of lead and the study of delirium in its sensitivity to unmeasured biases, but is closer to the study of delirium. If Wilcoxon's statistic is replaced by a better statistic, then the results are only slightly more insensitive, a maximum P-value of 0.044 at $\Gamma = 1.3$; moreover, at $\Gamma = 1.25$, the smallest possible estimated treatment effect is half a week longer duration of compensation.[34] From Table 9.1, a bias of magnitude $\Gamma = 1.25$ would be produced by an unobserved covariate, u_i, which doubled the odds of an injury claim in the after period and doubled the odds of a positive pair difference in time out of work.

In this example, an extremely small P-value derived from an assumption of ignorable treatment assignment was sensitive to biases that are not extremely large. Arguably, the original study was well designed, focusing on similar people before and after an intervention that was beyond their individual control, so perhaps extremely large biases are not anticipated here.

The Earthquake in Chile and Symptoms of Post-traumatic Stress

Chapter 6 discussed the study by Zubizarreta and colleagues of the possible effects of the powerful earthquake that struck Chile in 2010.[35] Table 6.3 described symptoms of post-traumatic stress in 2,520 matched pairs of people: a treated person in a region greatly shaken by the earthquake and a control from a region far removed from the earthquake. Recall that the scientific literature suggested that people exhibit discernably different symptoms in response to stresses that appear, to investigators, to be similar, and this pattern was evident also in Table 6.3.[36]

The earthquake example is the first opportunity to see a pattern or regularity in how sensitivity analyses turn out. In Chapter 10, such regularities will be examined in greater detail, but for now consider the earthquake and its effects. The literature said to expect heterogeneous reactions to the earthquake. It said that some people exposed to the earthquake would experience substantial symptoms, and others would exhibit few or negligible symptoms. As will be seen, if you select a test statistic designed to detect that pattern,

then you find the study is insensitive to much larger biases than you would have found had you blithely ignored the anticipated pattern when selecting a test statistic. Obviously, it is helpful that the literature was correct in its anticipations in this example.

Table 6.3 will be important, so reexamine its structure. On the upper left, there are 532 pairs in which both the treated and control individuals exhibited negligible symptoms of post-traumatic stress. In the fourth row and first column, there are 304 pairs in which the treated individual experienced high levels of symptoms but the control exhibited negligible symptoms. In the first row and fourth column, there are nine pairs in which the treated individual experienced negligible symptoms and the control exhibited substantial symptoms. So Table 6.3 has a diagonal, 532, 149, 30, and 6, representing the same symptom category for both individuals in a pair, and off-diagonals that are mirror images, such as 532 and 9. Another pair of mirror images is 569 in row two and column one, and 132 in row one and column two. There are six mirror image pairs in Table 6.3, plus four diagonal cells, making $(2 \times 6) + 4 = 16 = 4 \times 4$ cells of the table, without the totals.

Now, suppose that Fisher's hypothesis of no effect, H_0, is true, $r_{Ti} = r_{Ci}$ for every individual i, so the earthquake had no effect on symptoms. Fisher's hypothesis is just a hypothesis, and we will soon ask whether it resembles the data, but first let us be clear about what this hypothesis says and implies. If Fisher's hypothesis is true, then Harry's symptoms are Harry's symptoms, and changing Harry's exposure, Z_i, to the earthquake would not change his symptoms; that is, Harry is person $i = 17$ and if H_0 is true then $r_{T17} = r_{C17}$. The same is true of Sally. Perhaps Harry alternates between nightmares and insomnia, while Sally sleeps soundly, but if H_0 is true, then Harry would have his nightmares whether exposed to the earthquake or not, and Sally would sleep soundly whether exposed or not. Perhaps that is plausible, perhaps not, but it is what the hypothesis claims.

The study matched for gender, but ignoring this for a moment, if Harry and Sally were paired, and if Harry had a score of 80 or more and Sally had a score of 37 or less, then their one pair would be one of the $313 = 304 + 9$ pairs in the mirror image cells in (row, column) positions (4,1) and (1,4) in Table 6.3. Notice that, if H_0 is true, then the Harry-Sally pair is in one of these two cells regardless of which of them was exposed to the earthquake. If H_0 is true and Harry was exposed to the earthquake, their pair would be one of the 304 pairs in position (4,1); whereas if Sally was exposed, their

pair would be one of the nine pairs in position (1,4). If H_0 is true, and one person, Sally or Harry, was picked at random for exposure to the earthquake, then the Harry-Sally pair would be equally likely to show up in position (4,1) or (1,4), and the same is true of all 313 pairs in these two positions.

Indeed, the same is true of all of the pairs in mirror image cells. We can see that there is a substantial tension between three items: (i) the hypothesis H_0 of no effect of the earthquake, (ii) random exposure to the earthquake within each pair, and (iii) the data. After all, if (i) and (ii) were true, we expected the $313 = 304 + 9$ pairs to split evenly, 157.5 to 157.5, in cells (1,4) and (4,1), whereas the actual split is very lopsided, 304 to 9.

A simple test, but not a very good test, would ask this of Table 6.3: In how many pairs did the individual exposed to the earthquake have more severe symptoms? A little arithmetic gives the answer: in 1562 pairs the exposed individual had more severe symptoms, in 241 pairs the control had more severe symptoms, and the remaining 717 pairs fall on the diagonal of Table 6.3. Using McNemar's test and assuming random or ignorable treatment assignment gives a one-sided P-value of 5.3×10^{-237} testing no effect, and a two-sided P-value of twice that. A bias in treatment assignment of $\Gamma = 5.78$ would yield a one-sided P-value of at most 0.051. That is a high degree of insensitivity to unmeasured bias.

If, instead, we focused on the mini-study of just pairs in the mirror image cells in positions (1,2) and (2,1) of Table 6.3—that is, the split of 569 to 132—then McNemar's test indicates that much smaller biases could explain this pattern, with a P-value of at most 0.051 at $\Gamma = 3.67$. However, if we focused on the mini-study of just pairs in the mirror image cells in positions (1,4) and (4,1) of Table 6.3—the split of 304 to 9—then McNemar's indicates that much larger biases would need to be present to explain this pattern, with a P-value of at most 0.0023 at $\Gamma = 14$.[37] Remember, one of the major studies of heavy smoking and lung cancer becomes sensitive at $\Gamma = 6$, so the comparison of cells (1,4) and (4,1), or extreme versus negligible symptoms, is vastly more insensitive to bias than this smoking study.

In brief, if we focus on pairs in which one person had severe symptoms and the other had negligible symptoms, then Table 6.3 is extremely insensitive to unmeasured bias. Would something similar be seen if we had not grouped the data for tabular display in Table 6.3? Consider using the individual scores, from 34 to 170, measuring the symptoms of individual patients. Using Wilcoxon's signed rank test, the P-value is at most 0.050 for

$\Gamma = 7.15$. If we do not group scores, as in Table 6.3, but we just count the pairs in which the treated individual had more symptoms—the so-called sign test—then the P-value is at most 0.0492 for $\Gamma = 4.52$. So both of these standard tests say the study is sensitive to smaller biases than did the comparison of the (1,4) and (4,1) cells of Table 6.3.

In fact, Zubizarreta and colleagues used a test statistic S_m from Robert Stephenson's work that has optimal properties when a treatment only affects some people but has no effect on others.[38] Wilcoxon's test has some optimal properties if everyone is affected by the same amount, $r_{Ti} - r_{Ci} = \tau$ for every person i, but the literature suggests this will not happen with symptoms of post-traumatic stress. The literature suggests that $r_{Ti} - r_{Ci}$ will be large for some people and small or zero for many others. There is a version of the test statistic S_m for each positive integer m.

Consider, for example, four cases, namely, $m = 1$, $m = 2$, $m = 8$, and $m = 20$. The statistic S_m looks at m pairs at a time. Among those m pairs, it finds the largest absolute difference in symptom scores. For the one pair among the m pairs with the largest absolute difference in symptoms, it scores a 1 if the treated individual had more symptoms and 0 if the control had more symptoms. Finally, S_m sums the 1s and 0s over all sets of m pairs. If $m = 1$, then S_1 turns out to be the sign test used previously. If $m = 2$, then S_2 turns out to be Wilcoxon's signed rank test used previously. If $m = 20$, we look at 20 pairs, find the biggest difference in symptoms in these 20 pairs, and score a 1 or a 0 depending upon whether the treated or the control person in this pair had greater symptoms. So S_{20} cares very little about pairs in which both people had similar symptoms—such as the 532 pairs in the (1,1) cell of Table 6.3—but S_{20} cares a great deal about the $313 = 304 + 9$ pairs in the (4,1) and (1,4) cells of Table 6.3, where the difference in symptoms is large. In brief, S_8 and S_{20} use individual symptom scores, 34-to-170, but their focus is similar the McNemar test for the (4,1) and (1,4) cells of Table 6.3.

Using Stephenson's statistic S_8, the largest possible P-value testing the hypothesis of no effect is 0.024 for $\Gamma = 14.6$. Using S_{20} instead of S_8, the largest possible P-value testing the hypothesis of no effect is 0.023 for $\Gamma = 22.4$.

In summary, the choice of test statistic had a large effect on the reported degree of sensitivity to unmeasured bias in the earthquake example. The literature suggests that in response to a disaster some people will exhibit substantial symptoms of post-traumatic stress while others will exhibit neg-

ligible symptoms. A poor choice of test statistic says the effect is sensitive at about $\Gamma = 4.5$, much more sensitive to bias than one large study of smoking and lung cancer that is sensitive at $\Gamma = 6$, whereas a good choice of statistic says the effect of the earthquake is sensitive at $\Gamma = 22.4$. Statistical theory anticipates that S_8 will report greater insensitivity to unmeasured bias than Wilcoxon's statistic or the sign statistic when only some treated individuals respond to treatment.[39]

The earthquake example illustrates a general issue. Certain worldly situations, certain study designs, and certain methods of analysis are expected to be more insensitive to unmeasured biases than others. It would be helpful to know this when selecting the circumstances, the design, and the methods of analysis for an observational study. This topic is the focus of Chapter 10.

Taking Stock

Departures from randomized or ignorable treatment assignment resemble descending a gradual slope, not falling off a cliff. A sensitivity analysis asks, How far would we have to depart from randomized treatment assignment to alter the practical or qualitative conclusions of an observational study? Studies vary substantially in their sensitivity to unmeasured biases, and a sensitivity analysis locates a particular study in a well-defined spectrum of departures from random assignment. Specifically, a sensitivity analysis finds the location of a particular study in Table 9.1. The results of a sensitivity analysis are determined neither by the size of a P-value computed under the assumption of ignorable assignment, nor by the magnitude of an estimated treatment effect.

Design Sensitivity

What Is Design Sensitivity?

Can Studies Be Designed to Be Insensitive to Larger Biases?

In an observational study, an association between treatment received and outcome exhibited may not be an effect caused by the treatment, but rather some bias in the way treatments were assigned to individuals. How large would the bias—the departure from random assignment—have to be to alter the conclusions of the study? This question was answered in Chapter 9, and the answer was a number, a value of Γ in Table 9.1, indicating the smallest bias that could change the qualitative conclusions of an observational study. Having answered this question a few times in a few studies, we discover that some studies are insensitive to very large biases, while other studies are sensitive to small biases. Why is that? What makes some studies insensitive and others sensitive? Can we arrange or plan studies to ensure that they are insensitive? Can we anticipate the degree of sensitivity to bias from knowledge of the process that produced the data? Using that knowledge, can we identify and select better processes that produce data for observational studies? It is easier to hit a target if you know where the target is. It easier to hit a

bull's-eye if you can recognize a bull's-eye when you see one. What makes some studies insensitive to large biases, others sensitive to small ones?

Sensitivity Analysis and Design Sensitivity

A tool for answering these and related questions is the design sensitivity, $\tilde{\Gamma}$. Unlike the sensitivity parameter, Γ, which refers to data from an observational study, the design sensitivity, $\tilde{\Gamma}$, is a property of a research design. Here, a research design means a worldly situation, a way of collecting data in that situation, and a plan for statistical analysis. In thinking of natural experiments in Chapter 6, we tried to find situations in the world that were favorable for an observational study, perhaps with smaller unobserved biases and sharply defined interventions. In thinking of quasi-experimental devices in Chapter 8, we tried to augment the study's design with a view to reducing or eliminating particular ambiguities that might be mistaken for treatment effects. In the sensitivity analysis for the earthquake example in Chapter 9, we saw that different methods of analysis produce different evaluations of sensitivity to bias, and this is entwined with the nature of the treatment effect. The design sensitivity, $\tilde{\Gamma}$, is an aid to understanding how a particular worldly situation, a choice of data from that situation, a type of treatment effect, and a method of analysis combine to yield a level of sensitivity to unmeasured bias. If we understood these things, perhaps we could design and plan observational studies to be less sensitive to unmeasured biases. If we understood these things, perhaps we could create observational studies that are less likely to go wrong.

Design sensitivity, $\tilde{\Gamma}$, is computed under a probability model describing how the data are generated and analyzed. Under such a model, we know the true situation because we created the true situation. A probability model is similar to a laboratory. As in this laboratory, with a probability model generating the data, we know exactly what is happening because we created what is happening. Once we understand what happens in this laboratory, we can employ that understanding in the design of actual studies.

Design sensitivity, $\tilde{\Gamma}$, imagines that data came from a particular probability model, and then asks, If we had large quantities of data from this model, how sensitive would the conclusions be to unmeasured bias? If we performed the sensitivity analyses in Chapter 9 on data from a model that

we understand, how would things turn out? Somewhat more precisely, if we have P matched pairs from a particular model, and if we let P get very large, then what value of Γ would change the conclusions? That limiting value of Γ as $P \to \infty$ is the design sensitivity, $\tilde{\Gamma}$.

The Favorable Situation

Design sensitivity can be computed under (virtually) any model, but only some models are interesting and sensible. To use $\tilde{\Gamma}$ in a sensible way, we need to start with a situation in which we want to report insensitivity to unmeasured bias. If the study were truly very biased, the last thing we want to do is say it is not.

So we start with a situation in which there is a treatment effect and no unmeasured bias—call this the "favorable situation." In the earthquake example in Chapters 6 and 9, the favorable situation would mean that the matched comparison is fine as it is, and the earthquake caused some people to experience symptoms of post-traumatic stress. In the injury duration example in Chapters 8 and 9, the favorable situation would mean that the matched comparison is fine as it is, and the change in compensation caused some people to stay out of work longer. In any actual study, such as the earthquake and injury duration studies, no one knows whether we are in the favorable situation or not. The best we can hope to say in an actual observational study is that the conclusions are insensitive to moderately large biases. However, in the laboratory, we can create the favorable situation—a treatment effect without bias—and then see whether the results are sensitive to bias or not.

A Small Example of Design Sensitivity

Before using design sensitivity in the next section, consider its behavior and computation in the simplest nontrivial case. The goal, for now, is simply to make design sensitivity tangible.

Recall that in a matched pair study, Y_p is the treated-minus-control difference in outcomes. For example, in the study of lead in children's blood the differences Y_p were depicted in panel (a) of Figure 7.3, and for the study

of injury duration the differences Y_p for the affected "higher" wage group were depicted in Figure 8.3. In Chapter 9, the sensitivity of these comparisons was determined as the value of Γ that yielded a P-value above the conventional level of $\alpha = 0.05$. In both cases, we used Wilcoxon's signed rank statistic as an old, familiar, sturdy choice of test statistic, but the endnotes explored less familiar but better choices. In contrast, the design sensitivity, $\tilde{\Gamma}$, is not computed from an actual observational study but rather from a probability model that might generate data.

Suppose that the treated-minus-control pair differences in outcomes, Y_p, were unrelated observations from a Gaussian or Normal distribution with mean $\tau = 1/2$ and standard deviation 1. That is, we are in the favorable situation: there is a treatment effect, $\tau = 1/2$, with Normal errors, and there is no bias from unmeasured covariates. Because we cannot know when we are in the favorable situation, we would conduct a sensitivity analysis had we seen these data in an observational study. If we used Wilcoxon's signed rank statistic to test the hypothesis of no effect, then with sufficiently many pairs the results would be sensitive to a bias of $\tilde{\Gamma} = 3.171$. So this is a property of a particular Normal distribution, when analyzed using Wilcoxon's statistic, not a property of a specific observational study. The value, $\tilde{\Gamma} = 3.171$, is produced by a theoretical calculation that is not difficult to perform.

Suppose that we simulate a large data set from the situation just described, many pair differences, Y_p, from a Normal distribution with mean $\tau = 1/2$ and standard deviation 1. The Normal distribution with mean τ and standard deviation σ has variance σ^2 and is denoted $N(\tau, \sigma^2)$, so we will sample from the $N(1/2, 1)$ distribution. Figure 10.1 contains such a simulation with 100,000 observations. In Figure 10.1, the true mean is $1/2$, but the hypothesis of no treatment effect would imply, were it true, a mean of 0. So we are testing that the boxplot is centered at 0 when it is actually centered at $1/2$.

What does it mean that $\tilde{\Gamma} = 3.171$ when Wilcoxon's statistic is applied to data from the $N(1/2,1)$ distribution? If the sensitivity analysis in Chapter 9 is performed on the simulated data in Figure 10.1 using Wilcoxon's statistic, then for $\Gamma = 3$ the maximum one-sided P-value is 4.5×10^{-9}, whereas for $\Gamma = 3.3$ the maximum P-value is 1.00. That is, in this very large sample, as Γ moves from below $\tilde{\Gamma} = 3.171$ to above it, the maximum P-value jumps from near 0 to near 1. Closer in, at $\Gamma = 3.1$ the maximum P-value is 0.028, while at $\Gamma = 3.2$ the maximum P-value is 0.965. If we had drawn a million values of Y_p rather than 100,000, then the jump from 0 to 1 would occur very close

A Sample from N(1/2,1)

100,000 Simulated pair differences
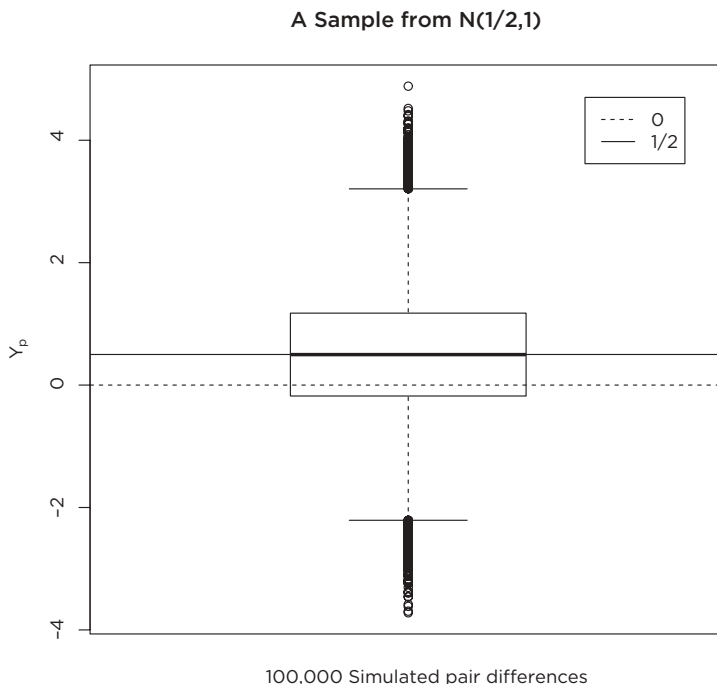
Figure 10.1. A simple sampling situation to illustrate the design sensitivity, $\tilde{\Gamma}$. There are 100,000 observations sampled from a Gaussian or Normal distribution with mean $1/2$ and standard deviation 1, or the N(1/2, 1) distribution. There are horizontal lines at the population mean, $1/2$, and the hypothesized mean of 0 for no treatment effect.

to $\tilde{\Gamma} = 3.171$. That is, $\tilde{\Gamma} = 3.171$ is the limiting sensitivity to bias as the sample size increases toward ∞.

In Chapter 9, in the context of the earthquake data, we considered Stephenson's statistic, S_8. For pair differences Y_p sampled from N(1/2, 1), like those in Figure 10.1, the design sensitivity for S_8 is $\tilde{\Gamma} = 6.5$, double the value for Wilcoxon's statistic. In other words, Stephenson's statistic, S_8, would report greater insensitivity to bias than Wilcoxon's statistic for the same data in Figure 10.1. For the data in Figure 10.1, at $\Gamma = 3.3$, Wilcoxon's statistic says the maximum *P*-value might round to 1.00, but Stephenson's statistic, S_8, says the maximum *P*-value rounds to 0.0000. At $\Gamma = 6 < 6.5 = \tilde{\Gamma}$, Stephenson's statistic, S_8, says the maximum *P*-value is 2.9×10^{-5}. Perhaps surprisingly, Stephenson's statistic, S_8, also has larger design sensitivity than the mean of the Y_p, usually considered the best statistic for Gaussian data. Specifically,

for the N(1/2, 1) distribution, the mean has design sensitivity $\tilde{\Gamma} = 3.5$, slightly better than Wilcoxon's 3.17, but much worse than 6.5 for S_8.

Stephenson's statistic, S_8, wins against Wilcoxon's statistic because the Normal distribution does not produce wild observations. For long-tailed distributions that do produce wild observations, such as the injury durations in Figure 8.3 before logs were taken, Stephenson's statistic S_8 is inferior to Wilcoxon's statistic. There are statistics that win against Wilcoxon's statistic for both long- and short-tailed distributions.[1]

What does a sensitivity analysis report when the data come from a theoretical model? In large samples, the answer is concisely summarized by the design sensitivity, $\tilde{\Gamma}$.

Heterogeneity and Causality

Sir Ronald Fisher and John Stuart Mill

In his *A System of Logic,* John Stuart Mill argued that causal inference entails comparing identical people under different treatments. Describing his "method of difference," Mill wrote: "If an instance in which the phenomenon . . . occurs and an instance in which it does not have every circumstance save one in common . . . [then] the circumstance [in] which alone the two instances differ is the . . . cause or a necessary part of the cause."[2]

In his *Design of Experiments,* in introducing randomized treatment assignment, Fisher attacked Mill's idea: "It is not sufficient remedy to insist that 'all the [units] must be exactly alike' in every respect except that to be tested. For this is a totally impossible requirement in our example, and equally in all other forms of experimentation."[3]

In the ProCESS Trial in Chapter 1, there was no way to find identical patients with sepsis, but it was straightforward to randomly assign heterogeneous patients to treatment or control. In Chapter 3, in testing Fisher's hypothesis of no treatment effect, we did not need identical patients; rather, we needed random assignment of treatments. In that sense, Fisher was right, and Mill was wrong.

But was Mill completely wrong?

Many experiments work hard to compare very similar units under alternative treatments. Laboratory experiments often use genetically engineered,

nearly identical strains of mice or rats—that is, mice or rats that are as homogeneous as possible. There are many observational studies that compare identical twins under alternative treatments.[4] In their study of the effects of the minimum-wage on employment, David Card and Alan Krueger compared Burger King restaurants in a state that had raised the minimum-wage to Burger Kings in a state that had not, knowing that few human creations are as homogeneous as Burger King restaurants.[5] Though Fisher won wide acceptance for randomized treatment assignment, Mill's idea of driving out heterogeneity is far from dead.

Unit Heterogeneity and Design Sensitivity

Was Mill completely wrong? If you cannot randomly assign treatments, does heterogeneity of experimental units affect design sensitivity? Are conclusions sensitive to smaller biases if the units under study are more heterogeneous? Are conclusions insensitive to larger biases if the units under study are more homogeneous? Laboratory scientists make fanatical efforts to eliminate noise, to remove heterogeneity, and in doing this they follow Mill's advice. So, again, was Mill completely wrong?

The question has two aspects. First, suppose one had a very large sample. Would less heterogeneity translate into greater insensitivity to bias? Second, suppose I can find less-heterogeneous units but only at the price of having fewer units. That is the situation with identical twins. Sibling pairs are far more numerous than identical twins, but they are also more heterogeneous than twins. Should I prefer fewer units with less heterogeneity or more units with more heterogeneity?

Both aspects have a theoretical answer, but let us start by making it tangible with a simulated example. Figure 10.2 offers you a choice of matched pair differences in outcomes, Y_p, from two observational studies that you might conduct. In parallel with Figure 10.1, the Y_p in Figure 10.2 were sampled from Gaussian or Normal distributions, with the same mean or treatment effect, $\tau = 1/2$. The boxplot on the left describes 400 differences Y_p from a Normal distribution with mean $\tau = 1/2$ and standard deviation 1, as in Figure 10.1. The boxplot on the right in the figure describes 100 differences Y_p from a Normal distribution with mean $\tau = 1/2$ and standard deviation 1/2; that is, fewer observations that are more stable, less heterogeneous. In

other words, Figure 10.2 compares a sample from the N(0,1) distribution to a sample from $N(0, 1/2^2)$ distribution. It is a tidy theoretical fact about the Normal distribution that the sample mean, $\overline{Y}$, behaves in the same way in the two boxplots. More precisely, the sample mean $\overline{Y}$ has a Normal distribution with expectation $\tau = 1/2$ and variance $1/400$ whether you have 400 observations from the Normal distribution with standard deviation 1 or 100 observations from the Normal distribution with standard deviation 1/2. So far as the sample mean $\overline{Y}$ is concerned, there is no reason to prefer one boxplot in Figure 10.2 to the other, because the loss of sample size is perfectly compensated by the reduction in heterogeneity. Nonetheless, the boxplots themselves look quite different.

A Choice of Two Studies
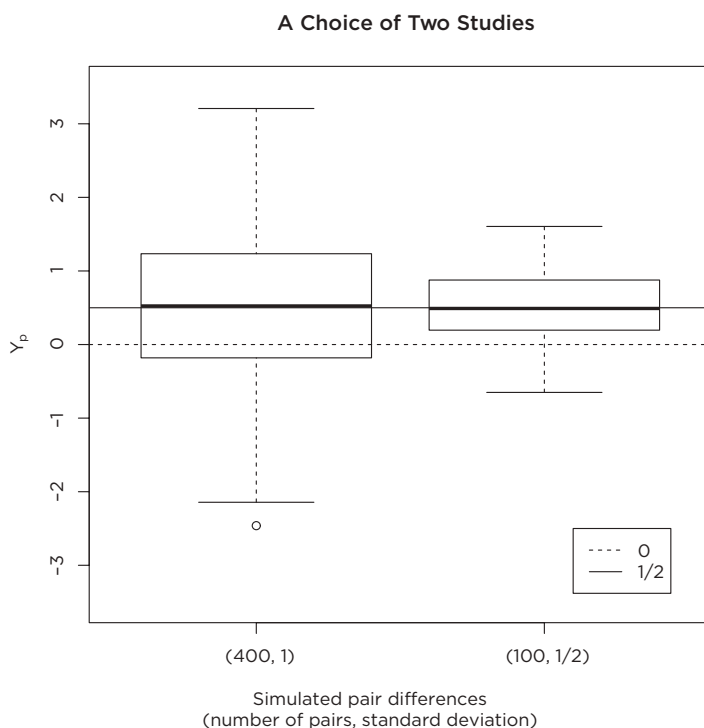
Figure 10.2. A choice of a larger but more heterogeneous study or a smaller but less heterogeneous study. The boxplot on the left depicts 400 observations sampled from a Normal distribution with mean $\tau = 1/2$ and standard deviation 1. The boxplot on the right depicts 100 observations sampled from a Normal distribution with mean $\tau = 1/2$ and standard deviation 1/2.

Mill says to prefer the boxplot on the right—fewer observations with less heterogeneity. Is Mill completely wrong?

Using the methods in Chapter 9, the boxplot on the left in Figure 10.2—more observations with more heterogeneity—becomes sensitive: (i) at $\Gamma = 2.6$ with a maximum one-sided P-value of 0.057 for Wilcoxon's statistic, (ii) at $\Gamma = 2.8$ with P-value 0.056 using the mean, $\overline{Y}$, and (iii) at $\Gamma = 4.2$ with P-value 0.058 using Stephenson's statistic, S_8. In contrast, the boxplot on the right in Figure 10.2—fewer observations with less heterogeneity—becomes sensitive (i) at $\Gamma = 5.1$ with a maximum P-value of 0.049 for both Wilcoxon's statistic and the mean, $\overline{Y}$, and (ii) at $\Gamma = 7.7$ with P-value 0.049 using Stephenson's statistic, S_8.

In short, Mill is not completely wrong. Heterogeneity directly affects causal inference, and less heterogeneity is better, consistent with Mill's advice and the practice of laboratory scientists. In Figure 10.2, the smaller but less heterogeneous study is more insensitive to unmeasured biases from an unobserved covariate u_i than the larger but more heterogeneous study. This occurs even though the sample mean, $\overline{Y}$, is equally accurate on the left and right of Figure 10.2, with heterogeneity and sample size offsetting one another. Is what happened in Figure 10.2 a peculiarity of this one simulated example? Straightforward calculations show that the pattern in Figure 10.2 is the expected pattern: reduced heterogeneity means greater insensitivity to bias when a loss of sample size just compensates for a reduction in heterogeneity.[6]

In large samples as well, reducing heterogeneity makes a study more insensitive to unmeasured biases. Consider the design sensitivities, the limiting sensitivity to bias as the number of pairs increases. For samples from the two Normal distributions in Figure 10.2, with mean $\tau = 1/2$ and standard deviations 1 or 1/2, the design sensitivities of Wilcoxon's statistic are $\tilde{\Gamma} = 3.171$ and $\tilde{\Gamma} = 11.715$, respectively. If the mean, $\overline{Y}$, is used instead of Wilcoxon's statistic, then the design sensitivities are $\tilde{\Gamma} = 3.528$ and $\tilde{\Gamma} = 13.003$, respectively.[7] The situation in large samples is more dramatic than in Figure 10.2 because we are no longer comparing 400 observations to 100 observations, and sample size is no longer an issue.

Mill was wrong in thinking we could find identical people under alternative treatments. He was right, however, in encouraging us to compare people who are similar in ways that strongly affect the heterogeneity of matched pair differences, Y_p.

How to Reduce Heterogeneity

Comparing identical twins is an obvious way to create pairs with reduced heterogeneity. There also are less obvious ways to do this.

A clever study by P. H. Wright and L. S. Robertson sought to determine the fixed road conditions that contribute to fatal collisions with roadside objects.[8] The difficulty is that most roads are comparatively safe affairs for sober drivers who are driving cautiously, in a safe car, in good weather, and with seatbelts fastened; however, no road is safe for an intoxicated driver, driving crazily, in an unsafe car, in foul weather, and with seatbelts unfastened. You would like to compare different roadside conditions with less heterogeneity from these other sources of risk. You would like to compare different road conditions for the same driver, in the same state of sobriety, in the same car, in the same weather, with seatbelts in the same state of use. Wright and Robertson did just that by comparing the sites of 300 fatal collisions with a roadside object to the sites of 300 nonaccidents. These 300 nonaccidents occurred with much fanfare a mile back on the road leading to the crash site, a location passed without incident minutes earlier. They discovered a substantial excess of crash sites with sharp curves on a downward slope. This is an example of Malcolm Maclure's case-crossover design.[9]

A different approach to reducing heterogeneity in pair differences in outcomes, Y_p, focuses on the construction of matched pairs. In a study of the effectiveness of for-profit and not-for-profit high schools in Chile, José R. Zubizarreta, Ricardo D. Paredes, and I proposed separating the process of matching from the process of pairing.[10] The outcome in this study was standardized test scores in 2006 while the students were in high school. First, the largest possible matched sample was constructed that balanced many covariates. Then, separately, students in this match were paired for a few variables thought to be highly predictive of the outcome. The goal was to balance many covariates but to reduce heterogeneity in paired differences in outcome, Y_p.

In this study of Chilean schools, the match picked 1,907 students attending for-profit high schools and picked 1,907 students attending not-for-profit high schools. This match balanced numerous covariates describing students before high school, including household income, mother's education, father's education, baseline standardized test scores in language, mathematics, and natural and social science, the number of books at home, and other

covariates. The baseline test scores were obtained in 2004, before these students attended high school. Once the $3,814 = 1,907 + 1,907$ students had been selected, they were paired to have similar baseline test scores before high school. The covariate balance was unchanged by the pairing, because covariate balance refers to the two groups of 1,907 students, not who was paired with whom.

To drive home the point, the process was done twice, pairing the same $3,814 = 1,907 + 1,907$ students in two different ways, something one would only do to illustrate a methodological point. One pairing used just baseline test scores, previously mentioned. The other pairing used all the covariates. Of course, the covariate balance is exactly the same with the two pairings, because the students are the same; the mean difference in outcomes, $\overline{Y}$, is also exactly the same, namely, $\overline{Y} = 17.5$ points higher test performance in 2006 for students in not-for-profit schools. However, pairing for baseline test scores alone made the pair differences, Y_p, less heterogeneous. The Y_p had standard deviation 90.9 when pairing for test scores, and standard deviation 105.5 when pairing for all covariates, a 16% difference; this is less dramatic than Figure 10.2, but not a trivial difference. In a sensitivity analysis for a test of no treatment effect, at $\Gamma = 1.6$, the upper bound on the P-value was 0.036 when pairing just for test scores, but it was 0.315 when pairing for all covariates.[11] This is a consequence of pairing the same students in two different ways. In brief, there can be a reduction in heterogeneity of Y_p, with a consequent increase in insensitivity to bias, by balancing many covariates while pairing only for covariates known in advance to be predictive of the outcome.

Subpopulations with Larger Treatment Effects

The magnitude of a treatment effect may vary from person to person. Other things being equal, a larger treatment effect is insensitive to larger unmeasured biases. If the study or the analysis can emphasize a subpopulation in which the treatment effect is larger, it may become clearer that an association between treatment and outcome in that subpopulation is actually an effect caused by the treatment, not a bias in the way treatments were assigned. Having established that the treatment has an actual treatment effect

in a subpopulation, it may become more plausible that weaker associations in other subpopulations are also effects caused by the treatment.

It may be that the treatment itself is heterogeneous, in dose or intensity, and perhaps intense versions of the treatment have larger effects. In Chapter 6, the study of bereavement by Lehman and colleagues focused on the sudden death in a car crash of a spouse aged 21 to 65 or a child under 18 living at home.[12] They focused on the sudden, senseless death at a premature age of someone close. Presumably they were aiming for an especially intense form of bereavement so that a long-term effect of bereavement, if it occurred, would be large, and if it did not occur, its absence would be striking.

Alternatively, there may be subpopulations defined by observed covariates, x_i, such that the treatment has a large effect in some subpopulations and a smaller effect in others. This is often called "effect modification"; it is an interaction between the treatment and a covariate. The Garki Project of the World Health Organization tested an intervention in Nigeria intended to control malaria. The intervention involved spraying with an insecticide, propoxur, and mass administration of a drug, sulfalene-pyrimethamine. The outcome was the density of malaria parasites in the blood. One analysis of the Garki Project found that the intervention had a substantial effect on young children and a much smaller effect on adults, with the consequence that the results for young children were insensitive to comparatively large biases, while the results for adults were sensitive to small biases.[13]

Finally, it may be that some treated people simply show little or no response to a treatment, while others respond strongly, but the investigators cannot identify before the fact who will respond and who will not. This pattern was expected in the study in Chapters 6 and 9 of the Chilean earthquake and subsequent symptoms of post-traumatic stress.[14]

The three patterns just noted can lead to larger design sensitivities and hence to greater insensitivity to unmeasured bias in particular studies. The strategy employed by Lehman and colleagues is a sound strategy: the primary analysis should compare an intense version of treatment to no treatment, excluding individuals exposed to mild or ambiguous versions of treatment. Strategies that discover effect modification can also produce studies insensitive to larger unmeasured biases. There are statistics designed to detect treatments that affect some people but not everyone, and, as seen in Chapter 9 with the Chilean earthquake, the use of these techniques in appropriate

contexts can result in conclusions insensitive to larger biases. Proof of these claims is slightly technical, so the interested reader should turn to the references.[15]

Can One Search for an Insensitive Finding?

Is it appropriate for an investigator to hunt for a finding that is insensitive to unmeasured bias? A careless hunt may produce nonsense because of the issue of multiplicity discussed in Chapter 8. If you test many hypotheses and do nothing about multiplicity, then you are sure to make many mistakes, repeatedly and frequently finding something where there is nothing. However, it is not difficult to control error rates when testing many hypotheses, and the cited technical references discuss the issue.

There is, however, a simple and safe way to hunt for a finding insensitive to unmeasured bias. The technique is most useful in large studies, often studies from administrative data systems, such as Medicare in the United States or the national educational testing program Sistema de Medición de Calidad de la Educación (SIMCE) in Chile, because in these cases the sample size is often much larger than is needed. The technique randomly splits the available data into two parts, 10% and 90%: a planning sample and an analysis sample. The investigator inspects the 10% planning sample, and on that basis develops a plan for a primary analysis. The investigator then discards the planning sample and carries out the planned, limited primary analysis on the unexamined 90% analysis sample. In this way, the investigator performs two independent studies: a small, speculative pilot study, and a large, thoughtfully planned, narrowly focused confirmatory study. In a large study, the loss of 10% of the sample often has negligible consequences, and there can be large gains from having a better plan for one primary analysis.[16]

✳ Austin Bradford Hill and Doses of Treatment

In 1965, Sir Austin Bradford Hill wrote an influential essay about distinguishing causal effects from biases. The original essay is thoughtful, wise, and very much worth reading, but the essay's legacy is a different matter. Hill asked, "Our observations reveal an association between two variables,

perfectly clear-cut and beyond what we would care to attribute to the play of chance. What aspects of that association should we especially consider before deciding that the most likely interpretation of it is causation?"[17] The tone of this question is the tone of the entire essay. The essay speaks repeatedly of considerations, weighed thoughtfully, in forming a reasoned judgment for which the investigator assumes full responsibility. What you make of a consideration in forming a judgment is your responsibility.

Subsequent literature often, perhaps typically, demoted Hill's considerations into "criteria for causality," a checklist mechanically applied. Such a checklist is an external standard against which one compares one's own study, blaming the author of the checklist, not oneself, if a mistaken conclusion is produced while maintaining conformity with the checklist. The mentality of a checklist is exactly opposed to the thoughtful, responsible tone of the original essay.[18]

About doses of treatment, Hill wrote:

> Biological gradient: Fifthly, if the association is one which can reveal a biological gradient, or dose-response curve, then we should look most carefully for such evidence. For instance, the fact that the death rate from cancer of the lung rises linearly with the number of cigarettes smoked daily, adds a very great deal to the simpler evidence that cigarette smokers have a higher death rate than non-smokers. That comparison would be weakened, though not necessarily destroyed, if it depended upon, say, a much heavier death rate in light smokers and a lower rate in heavier smokers. We should then need to envisage some much more complex relationship to satisfy the cause-and-effect hypothesis.[19]

Here, Hill says that if the relationship between lung cancer and the amount smoked had had an inverted *U*-shape—a rise followed by a decline—rather than its actual steady rise, then we would be forced to ponder the origin of this peculiar shape.

At odds with Hill's actual expressed view, Hill's statement about dose–response is often understood to require a "test for trend" computed under the assumption that treatment assignment is ignorable. Tests for trend, such as Mantel's test for trend, are designed to boost power in the absence of bias; they do not distinguish bias from treatment effects.[20] In contrast, as quoted here, Hill assumed at the outset of his essay that chance had been eliminated

as an explanation, so a boost in power is not relevant; his essay asked whether the observed association was produced by bias or causation.

If there is a steady increase in the size of the treatment effect with increased doses of treatment—a big *if*—then calculations based on design sensitivity indicate that, in large samples, comparison of a high-dose group with a zero-dose control group will be insensitive to larger biases than analyses that include minor doses.[21] This is the structure of the study of bereavement by Lehman and colleagues, not the structure of a test for trend.

It is common, if not routine, that observational studies first show that high doses of a toxin cause great harm: only enormous biases from unmeasured covariates could explain such a strong association as the one seen when comparing high doses to zero-dose control. Then attention turns to a weaker and more sensitive association between lower doses and more limited and erratic harms. These weaker associations may become plausible as an effect caused by the toxin in part because we are convinced that the toxin does cause harm at high doses. See, for instance, the discussion of fetal alcohol syndrome in Chapter 7 and the use, with heavy smoking, of the inequality of Cornfield and colleagues in Chapter 9. Additionally, the study of high doses may guide us about the manner in which the toxin produces harm—the intermediate steps, symptoms, and precursors—and these intermediate forms may be evident at lower doses. None of this has to do with boosting power using a test for trend under the assumption of ignorable treatment assignment.

∗ Clustered Treatment Assignments and Sensitivity to Bias

∗ Randomly Assigning Treatments to Clusters of Individuals

Some randomized experiments do not assign treatments to individuals but rather to clusters or groups of individuals so that everyone in the same cluster receives the same treatment.[22] As mentioned in Chapter 5, this was true in the PROSPECT randomized trial, a study of the treatment of depression by primary care physicians.[23] The PROSPECT study paired 20 primary care medical practices to form 10 pairs of two similar practices. A fair coin was

flipped 10 times, and one practice in each pair was given a depression-care manager, a specialist nurse, while the other practice in the pair served as the control. It was hoped that the presence of the depression-care manager would improve the diagnosis and treatment of depression throughout the practice. Every patient in the same practice was randomized to the same treatment; either there was a depression-care manager at the practice or there was not. Here, there were 20 clusters of patients, even though more than 8,000 patients were screened for depression at the 20 practices.

Randomizing clusters does not create bias in a randomized trial, provided that appropriate analyses are performed. However, a study that randomizes clusters of individuals can be a little noisier and less precise than a study that randomizes the same number of people as individuals. This loss of precision would be larger if depression outcomes tended to be similar within practices and different among practices receiving the same treatment, and it would be smaller if patients and outcomes were about the same in all practices. Had patients been randomly assigned to clusters of equal size, and clusters randomized to treatment or control, then there would be almost no difference between randomization of clusters and randomization of individuals. The loss of precision occurs because the clusters occur naturally, and people are not randomly assigned to their clusters.

In brief, in experiments, you randomize clusters rather than individuals only because the nature of the treatment requires this. The situation is different in observational studies. Can you see why?

* Observational Studies in Which Treatments Are Assigned to Clusters of Individuals

Some observational studies seem analogous to experiments with treatments assigned to clusters of individuals. For instance, Carlo Del Ninno and colleagues studied the effects of the severe floods that struck some villages in Bangladesh in 1998 and spared others.[24] In particular, they studied illness among children, including the mean number of days of illness among children in a village in the two weeks after the flood. A flood can injure people directly, affect supplies of fresh water and food, damage homes, or overwhelm medical resources, and each of these can affect the health of children.

Floods do not select villages at random for flooding, so there is, at best, a limited analogy with clustered randomization. Is this observational study most similar to a clustered randomized experiment, not an experiment randomized at the individual level? Young Sally lives in a flooded village while her young cousin Harry lives in an unflooded village. Their mothers are sisters who chose to live in different villages. Is it natural to think of the flood striking Harry's village and sparing Sally's village? Or is it more natural to think of Harry and Sally swapping villages, the two sisters reversing their choices about where to live? Randomization would swap individuals one at a time; clustered randomization would swap whole villages. Because this is an observational study, not an experiment, and because no randomization took place, it may be unclear what constitutes an analogous experiment. Because randomization was not used, the treatment assignments in the observational study could be biased in ways that an analogous experiment would not be biased, regardless of which, if either, experiment was thought to be analogous.

In this setting, there is a useful fact about design sensitivity. Remember, design sensitivity refers only to large samples, many villages, each village having at least a few children; so the situation would be more complicated in small samples. Here is the useful fact. Consider a large sample of many villages, with each village having at least a few children. Suppose additionally that children in the same village may be similar, but they are like children everywhere in that they are not absolutely identical in their health.[25] Then the clustered study will have a larger design sensitivity than the individually assigned study: it would take larger biases to explain away the same treatment effect in the clustered study.[26]

This is a technical result, and there are quite a few details and caveats not discussed here, but let me offer some intuition in support of the technical result. In the clustered study, biases must push whole villages toward treatment or control; that is, the biases cannot push the healthiest individual children into flooded villages. In contrast, if Harry's mother chose to live in a dry climate far from the river because Harry was a sickly child, then that bias operates on individual children. The technical result says that a bias that affects clusters—whole villages—must be larger to do the same damage as a bias that affects individuals.

If you control treatment assignments, then you want to make the most of your control and randomize individuals. If nature controls treatment as-

signments and you want to limit the damage nature can do to your study, then you want nature to have fewer choices; that is, you want nature to be forced to assign whole clusters to the same treatment. As in chess, it can be helpful if your moves are fairly unconstrained, and helpful if your opponent's moves are very constrained. In an observational study, nature controls treatment assignments, so it is helpful to you if nature faces more constraints.

Some recent work has sought to optimize the construction of matched, clustered observational studies; this is so-called multilevel matching.[27] In the examples in this work, the clusters are schools that may be assigned to treatment or control. The algorithm entertains every possible match of students in every treated and every control school; then it selects from these the best possible match of treated and control schools. If there are 100 treated schools and 200 control schools, then $100 \times 200 = 20,000$ matched samples are constructed: each of 100 treated schools with each of 200 potential control schools. From these 20,000 matched samples, the best 100 nonoverlapping matched samples are retained, yielding 100 pairs of a treated school and a control school. The algorithm finds similar students in similar schools under treatment and control.

Taking Stock

Design sensitivity anticipates the outcome of a sensitivity analysis. It is useful in anticipating the circumstances that would produce a study insensitive to moderately large unmeasured biases. Design sensitivity is computed under a model for the generation of the population, under a specific design for observing a large sample from the population, and under specific methods of analysis. If we know which populations, designs, and methods of analysis would lead to insensitivity to bias, then we can make informed choices when planning an observational study.[28]

Matching Techniques

Concepts and Issues

The Situation So Far

For John Stuart Mill, causal inference compared identical people under alternative treatments. Table 5.6 shows that Mill's idea is not a possible solution; indeed, it is neither possible nor a solution. It is not possible because, with 30 measured covariates each at three levels, there are vastly more types of people than there are people on earth. It is not a solution because adjustments for measured covariates do not ensure comparability in terms of covariates that were not measured. Mill's idea is not completely wrong—driving out heterogeneity improves design sensitivity, as seen in Chapter 10—but it is not realistic to compare identical people under alternative treatments.

Identical people do not exist, but comparing identical people is not necessary for causal inference. In randomized experiments in Chapter 3, for causal inference it was sufficient that treatments were assigned at random, by flipping a fair coin, so that the treatment a person received was unrelated to every attribute of that person. In Chapter 5, we saw that if treatment assignment is ignorable given observed covariates x—a big if—then it suffices to compare two people with the same propensity score, λ_x, that is, two

people who are the same on a single covariate. In this case, the problem in Table 5.6 does not arise because there is just one covariate. A simple approach to matching in observational studies estimates the propensity score, λ_x, forms matched treated–control pairs with very similar estimated propensity scores, checks that matching has indeed balanced the observed covariates, x, and then turns attention to the central and larger task of addressing bias from unmeasured covariates, u. This central and larger task speaks to the possibility that treatment assignment is not ignorable given x, so adjustments for x do not suffice to remove bias in estimated treatment effects.[1]

Forming pairs matched for the propensity score, λ_x, is a practical approach, but it is possible to use modern methods to do more. What more might one wish to do? Here are a few things. (i) Two people with the same propensity score, λ_x, may be very different in terms of x. Although it is typically impossible to find people who are identical in terms of all of x, it is often possible to find people who have similar propensity scores, λ_x, and additionally similar values of a few important covariates from x. In contrast, matching for the propensity score alone does nothing to ensure close matches on important aspects of x. (ii) Propensity scores use probability to balance the distribution of observed covariates, x, in treated and control groups. It is often possible to force closer balance on some aspects of x—to avoid leaving the matter to chance—and it may be necessary to do this when people are thinly spread over many categories of a nominal covariate, such as many types of surgical procedures. (iii) Pair matching—one treated individual with one control—may not make the most effective use of the available data, and more flexible structures may use more data more effectively.

Surgical Outcomes and Costs at Hospitals with Superior Nursing

A recent matched study by Jeffrey Silber and colleagues compared outcomes and costs of general surgery at 35 hospitals with superior nursing environments, called focal hospitals, and at 293 control hospitals.[2] The hospitals were in the states of Illinois, New York, and Texas, and the data described Medicare patients. The 35 focal hospitals with superior nurse environments were defined by two attributes: (i) they had been labeled by the American Nurses

Credentialing Center as magnet hospitals, and (ii) they had a nurse-to-bed ratio of one or more. The control hospitals had neither of these two attributes.[3] The causal effect or counterfactual is: Would a patient, say, Harry, have better or worse outcomes and costs if his surgery were performed at one of the focal hospitals rather than at one of the control hospitals?[4]

The study formed 25,072 pairs of two Medicare patients undergoing general surgery, one at a focal hospital and the other at a control hospital. Within each pair, the two patients underwent the same surgical procedure as recorded in four-digit principal procedure codes; there were 130 different surgical procedures. Additionally, the matching balanced 42 other patient covariates, including age, year of admission, sex, race, emergency admission status, transfer-in status, propensity score, risk-of-death score, and 31 comorbidities such as congestive heart failure or a prior heart attack, making a total of $172 = 130 + 42$ observed covariates, x_i. The article and its online appendix documented the comparability of the focal and control groups in terms of each of the 172 observed covariates x_i and conducted a sensitivity analysis for bias from an unobserved covariate u_i. In particular, the focal-minus-control difference in means in matched pairs for each of the 172 covariates was less than 1/20 of the standard deviation of the covariate before matching. After matching, the 172 covariates looked similar in the focal and control groups.

The 30-day mortality rate was materially lower at focal hospitals: 4.8% for focal hospitals and 5.8% for control hospitals.[5] Control hospitals made much more extensive use of the intensive care unit, a costly resource, and this was true even for patients whose baseline risk of death was low.[6] Determining costs and payments is not straightforward for Medicare patients, but several attempts to do this suggest that focal hospitals do not cost much more than control hospitals, and may well cost less by virtue of reducing the consumption of expensive resources such as the intensive care unit.[7]

Although identical people do not exist, it is possible to form treated and control groups that balance 172 observed covariates. The match was constructed using various techniques described in this chapter, including (i) exact matches for 130 surgical procedures, (ii) calipers on propensity scores and risk scores, and (iii) an algorithm that minimized the total within-pair distance on all observed covariates.[8] Matching was part of the design of the study: it was completed before outcomes were examined.

Presentation of a Matched Comparison in a Scientific Article

> No theory is kind to us that cheats us of seeing.
>
> —HENRY JAMES[9]

Competent presentation of a matched comparison in a scientific article demonstrates that matching has been effective in forming treated and control groups with similar distributions of the observed covariates, x_i. This demonstration may consist of one simple table, similar to Table 1.1 for a randomized trial, showing that treated and control groups are similar in terms of x_i after matching; see, for instance, Tables 8.1, 12.2, and 12.3. Additional tables or graphs may present additional detail, going beyond a mean or a proportion, as in Table 6.1 or Figure 8.1. One common approach presents, in the body of a published article, a single table describing the balance attained for many important observed covariates, x_i, and supports this, as needed, with additional detail in an online supplemental appendix.[10] The additional detail may describe less important covariates; for instance, when matching for surgical procedures, common procedures might be shown to be balanced in the published table, with many rare procedures shown to be balanced in an online supplement. Alternatively, the additional detail might describe the distribution of a continuous covariate, such as age or education in Table 6.1, where the published article reports only a median or a mean. The adjustment performed by matching is transparent and open to view: aided by an online appendix, the reader may examine the matched comparison in detail. Many readers will trust that the investigator has been careful and so will not examine all this detail, but all readers know that critical discussion of the research is informed by full access to details of the matched comparison.

Typically, matching excludes some controls.[11] In a table documenting the comparability of matched treated and control groups, it may be useful to show the manner and degree to which matching has altered the control population. If the matching has excluded some treated individuals—so-called subset matching—then it may be useful for this same table to show how matching has altered both the treated and control populations.[12]

A matched comparison should include one primary analysis, selected during the design of the study before any outcomes were examined.[13] In this

sense, a matched observational study resembles a randomized trial. With appropriate planning and with appropriate adjustments for multiple inferences, this primary comparison may involve several primary outcomes or comparisons, as is sometimes done in the protocol for a randomized trial.[14] This primary analysis uses matched pairs or matched sets in a simple and direct way. Although many published articles include additional analyses labeled as secondary and exploratory analyses, it would be a source of concern if the study's main conclusions were at odds with the findings of its one prespecified, primary analysis. An investigator who performs many complex analyses exercises enormous choice in the presentation of scientific findings, and a simple, prespecified primary analysis is intended to place a sharp limit on this sort of spin-doctoring.

Consistent with the quoted remark of Donald Rubin in Chapter 9, the presentation of a matched comparison typically affirms that the construction of the matched sample was completed before the examination of outcomes, "thereby assuring the objectivity of the design." A truthful affirmation permits the investigator to work at improving covariate balance, rejecting an initial match as inadequate because a covariate is not balanced. Because covariate balance is improved in ignorance of the study's outcomes, the processes of improving covariate balance cannot be used to spin the study's conclusions in a desired direction.

Matching and the Grounds upon Which Observational Studies Are Disputed

Matching may use technical tools to balance many observed covariates, x, but it leaves in its wake a simple structure, perhaps matched pairs, in which treated and control groups are readily seen to be comparable in terms of each measured covariate. With concerns about the measured covariates removed from the picture, our attention turns to the challenging issues that determine whether or not an observational study is convincing.

Matching is not simply, perhaps not primarily, about the observed covariates, x. Matching creates the structure and sets up the comparisons that frame the study and its subsequent critical evaluation. Discussions of biases from unobserved covariates, u, take place within this structure. If the struc-

ture facilitates discussion of unobserved biases, then the discussion may be illuminating, perhaps compelling.

Aspects of matching for observed covariates affect the discussion of unobserved covariates. In Chapter 10, we saw that balancing many covariates and pairing for a few highly predictive covariates can reduce heterogeneity of matched pair differences, Y_i, thereby making the study insensitive to larger biases. Also in Chapter 10 we saw that the discovery of effect modifiers can make the study insensitive to larger biases; moreover, certain types of matching facilitate the discovery of effect modifiers.[15] In Chapters 7 and 8, the construction of multiple control groups, evidence factors, and comparisons involving counterparts can be facilitated by matching.[16] Again, matching eliminates observed covariates from some comparisons, thereby simplifying these comparisons; see, for instance, Figures 7.2–7.3 and 8.1–8.3.

In contrast, when investigators use analytical methods to adjust for observed covariates x, it can be difficult to see whether the investigator has compared people who are comparable in terms of x, and who has been compared with whom.[17] A matched comparison offers for close inspection groups visibly comparable in terms of x. In contrast, analytical methods may offer an estimate of the treatment effect whose only justification is that certain procedures were used. There is no visible demonstration that the adjustments for x have succeeded. It can feel like buying a house without seeing it, being offered instead an audio recording the sounds of hammering and sawing. With so much fine hammering and sawing like that, how could the house be less than perfect?

Ornate adjustments for observed covariates can, and often do, inhibit critical discussion. Is inhibiting critical discussion a good thing? Should we be impressed or concerned when critical discussion is limited? To be compelling, an observational study must speak to the issues that make observational evidence debatable; it must engage, not avoid, the debate. As Mill put it, "He who knows only his own side of the case, knows little of that."[18]

A compelling observational study is one that has received the implicit endorsement of surviving critical discussion largely unscathed. We build conviction that a theory T is true if we have reasonably direct access to substantial, clear evidence that supports T, if no substantial, clear evidence undermines T, if the evidence under consideration is free of rectifiable omissions, and if critics of T have offered counterarguments that are vague, weak, or self-serving.[19]

∗ **Techniques for Constructing a Matched Comparison**

∗ Covariate Distances

Most modern methods select a matched sample that is optimal in the specific sense of minimizing the total distance within matched pairs. If person i is treated and person j is a control, the distance between person i and person j is a measure of the difference between their observed covariates, x_i and x_j. If x were a single covariate, say age, then a reasonable distance is the absolute difference in age, $|x_i - x_j|$, or perhaps the squared difference, $(x_i - x_j)^2$. Typically there are many covariates, not one, and the distance is chosen to set priorities among covariates. Some priorities are set by statistical theory, and others by relative importance of different covariates in a particular scientific field.

Because matching for the propensity score alone, λ_x or $\lambda(x)$, can balance all the covariates on its own, the distance usually emphasizes differences on the propensity score. A common strategy places a caliper, $\kappa > 0$, on the propensity score, requiring $|\lambda(x_i) - \lambda(x_j)| \le \kappa$ if i and j are to be matched. This is sometimes implemented by setting the distance between i and j to ∞ if $|\lambda(x_i) - \lambda(x_j)| > \kappa$.[20]

To match for both height and weight, something needs to be done to say whether an inch of height counts more or less than a pound of weight. Additionally, two covariates may be so similar that they are barely distinct; for instance, this might or might not be true of two measures of obesity, such as the body mass index and the sagittal abdominal diameter. When should two related measures count as two covariates and when should they count as one? Are there intermediate possibilities, where the second covariate only contributes what is not already obvious from the first? These questions are addressed by a distance proposed by an eminent statistician from India named Prasanta Chandra Mahalanobis. Small adjustments are commonly made to the definition of the Mahalanobis distance so that it is not thrown off by one or two peculiar observations or by rare binary variables, both of which commonly occur in matching.[21]

So a simple distance places a caliper on the propensity score; when persons i and j are within the caliper on the propensity score, it defines the distance between i and j to be a robust version of the Mahalanobis distance.[22] It is usually wise to let the propensity score balance many covariates but to

define the Mahalanobis distance using the most important covariates. Again, this wisdom derives from the issue in Table 5.6: identical people do not exist, so the investigator who wants paired individuals to be close must set some priorities among covariates, balancing many but pairing for a few. If another study or an external data set offers a score that predicts the outcome—say, a score predicting risk of death—this external prognostic score may be a priority for inclusion in the Mahalanobis distance.[23]

If the investigator plans to ask whether a particular covariate is an effect modifier, it is convenient that all or most pairs are exactly matched for this covariate. This is done by adding to the distance a large number or penalty whenever persons i and j differ with respect to this covariate. This is called "near exact" matching because, for a sufficiently large penalty, it maximizes the number of pairs exactly matched for this covariate but tolerates a mismatch when it cannot be avoided. Seeking exact or near exact matching makes sense only for a few covariates; see, again, Table 5.6 where exact matching for many covariates is not possible.[24]

✳ Minimizing the Total Distance

Once distances are defined, the treated and control individuals are allocated to nonoverlapping pairs. An optimal algorithm minimizes the sum of the distances for paired individuals. In Table 11.1, there are six people: three treated people numbered 1, 2, 3, and three controls numbered 4, 5, 6. The entries in the table are the distances between people. For instance, control 4 is close to treated person 1 with a distance of 0.1, is farther from treated person 3 with a distance of 10, and is at an infinite distance from treated person 2, perhaps because a caliper on the propensity score has been violated.

Again, the problem is to pair individuals to minimize the total distance while ensuring the pairs do not overlap. In Table 11.1, it is easy to see that the optimal pairing is (1,4) with distance 0.1, (2,5) with distance 0.2, and (3,6) with distance 0.1, yielding a total distance of $0.1 + 0.2 + 0.1 = 0.4$. Large matching problems with thousands of people are less obvious, but fast algorithms that solve large problems have existed since 1955.[25]

An intuitive but mistaken alternative to optimal matching uses a greedy algorithm. A greedy algorithm solves a multistep problem by making the best choice at the first step, never reconsidering that first choice, making the

Table 11.1. Hypothetical distance matrix based on covariates between three treated
               individuals (1, 2, 3) and three controls (4, 5, 6)

	Control i = 4	*Control i = 5*	*Control i = 6*
Treated $i = 1$	0.1	0.0	100.0
Treated $i = 2$	∞	0.2	5.0
Treated $i = 3$	10.0	5.0	0.1

best choice among the remaining choices at the second step, never reconsid-
ering, and so on. Greedy algorithms provide optimal solutions to a few prob-
lems but not to the matching problem. It is easy to see why in Table 11.1. The
one best match in Table 11.1 is treated person $i = 1$ with control $i = 5$ for a
distance of 0, and the greedy algorithm grabs it. Because it cannot recon-
sider, the greedy algorithm never recovers from its mistaken first move. The
greedy algorithm next pairs treated $i = 3$ to control $i = 6$ for a distance of
0.1, then is forced to pair treated $i = 2$ to control $i = 4$ at a distance of ∞, for
a total distance of $0 + 0.1 + ∞$. In Table 11.1, there are six possible pairings,
and four of the six avoid the infinite distance, so the greedy algorithm has
produced one of the worst pairings. When the greedy algorithm performs
poorly, it typically follows the pattern in Table 11.1, with early choices that
look quite good, discovering near the end that it has painted itself into a
corner and must accept some very poorly matched pairs. A greedy algorithm
may perform well when there are many potential controls available, but an
optimal algorithm is a safe bet.

Recall from Chapter 10 that close pairs may reduce heterogeneity in the
matched pair difference in outcomes, Y_i, with the consequence that conclu-
sions are insensitive to larger biases, Γ. In light of this, there are good reasons
to prefer pairs closely matched for important covariates. An optimal matching
algorithm achieves this goal to the extent that it is possible to do so.

*   Fine Balance and Related Techniques

Fine balance means forcing a covariate to be balanced without worrying
about who is paired to whom.[26] To finely balance gender means that every
treated man matched to a control woman is counterbalanced by another pair
in which a treated woman is matched to a control man; therefore, treated

and control groups have exactly the same number of women and exactly the same number of men.

Propensity scores use luck to balance covariates, whereas fine balance means forcing balance. Luck is reasonably reliable in large studies for common events. If the study is large with many men and many women, it is likely that propensity scores will do a reasonable job of balancing gender. Fine balance is most useful when people are thinly spread through many categories of covariate. For instance, in a study of surgical outcomes, some surgical procedures are common, but many others are rare. Use of probability may not balance many rare types of surgery because within each type there are only a handful of patients. Flip a fair coin three times, and you might get three heads, a perfect imbalance. That will not happen if you flip a fair coin 300 times. To use the law of large numbers to balance a category of a covariate, you need numbers that are, well, at least not small.[27]

Sometimes fine balance is not achievable. If there are 30 women in the treated group and if all treated people will be matched but there are only 29 women in the group of potential controls from which pairs will be drawn, then fine balance is not achievable for gender. Near-fine balance means coming as close to fine balance as the data will allow.[28] In the previous illustration, near-fine balance means ensuring that all 29 control women are indeed matched, so the imbalance, 30-to-29, is as small as possible. There may be many matched samples that use all 29 control women, and some may be better than others in terms of balancing another variable, say, the binary covariate unemployed. Refined balance means coming as close as possible to fine balance for certain covariates such as gender, then among all matches that do that, picking a match that comes as close as possible to fine balance for certain additional covariates, say, unemployed. Refined balance can have more than two stages. A recent application of refined balance had six stages, 2.8 million categories in stage six, and in every stage the match exhibited better balance on covariates than did the most balanced of 10,000 randomized experiments.[29] That is, the balance was much better than luck can be expected to produce for observed covariates.

Technically, the various forms of fine balance are constraints imposed on the problem of matching to minimize the total distance, as discussed in the previous section. That is, the total distance within matched pairs is minimized subject to the added requirement that fine balance is achieved.

*   Matching Each Treated Individual to K Controls

When many potential controls are available, a simple alternative to matched pairs is to match each treated subject to K controls for some number $K \geq 2$.[30] For instance, each treated subject might be matched to $K = 2$ controls.

   Methods for pair matching in previous sections carry over without change to matching with K controls. One may match to minimize the distance between each treated subject and its K matched controls, summed over all matched sets, perhaps imposing fine balance constraints.

   There is one argument that favors matched pairs, and there are two arguments that favor matching with K controls. Using optimal matching in a given data set, matched pairs will be more closely matched for observed covariates, x, because for pairs the algorithm sets aside inferior controls that it is forced to use when matching $K = 2$ controls to each treated subject—this is the argument that favors matched pairs.

   The first argument that favors using $K \geq 2$ matched controls refers to the increase in sample size. In a restricted but useful way, matching with K controls increases the precision of estimates of treatment effects when the treatment assignment is ignorable. Under one very simple model for continuous responses, matching each treated subject to K controls produces an estimate of the average treatment effect with a variance proportional to $1 + 1/K$.[31] Table 11.2 shows how the variance changes from matched pairs, $K = 1$, to matching with infinitely many controls matched to each treated subject, $K = \infty$. Adding controls stabilizes the control group, not the treated group, so the estimate with infinitely many controls remains unstable because the number of treated subjects has not increased. In Table 11.2, using two controls, $K = 2$, rather than pairs, $K = 1$, reduces the variance multiplier from $2 = 1 + 1/1$ for pairs to $1.5 = 1 + 1/2$ for triples, and this is halfway to the limit of 1 for infinitely many controls, $K = \infty$. Use of $K = 4$ controls rather than $K = 2$ controls cuts the remaining distance to $K = \infty$ in half again, from 1.5 to 1.25. The difference between using $K = 10$ controls and $K = 50$ controls is barely noticeable in Table 11.2. This calculation suggests that use of $K = 2$ controls rather than pairs can confer large benefits at small costs, but the opposite is true of the use of $K = 50$ controls rather than $K = 10$ controls.

   The second argument that favors using $K \geq 2$ matched controls no longer assumes treatment assignment is ignorable. Rather, the second argument is

Table 11.2. Relative size, $1 + 1/K$, of the variance of the estimate of the average
treatment effect with K controls matched to each treated, under a simple
model with constant variances

Number K of controls	1	2	4	10	20	50	∞
Relative variance $1 + 1/K$	2	1.5	1.25	1.1	1.05	1.02	1

concerned with sensitivity to unmeasured bias in Chapter 9 and design sensitivity in Chapter 10. We need to distinguish a binary response, say, dead or alive, from a continuous response, say, blood lead levels in Chapter 7. With a continuous response, use of $K \geq 2$ matched controls brings about a modest increase in the design sensitivity, $\tilde{\Gamma}$. This does not occur if the response is binary.

A precise statement and proof is somewhat technical.[32] Here is some intuition in place of a proof. It is often said that you cannot fit a square peg in a round hole, but that is not true. You can fit a big round peg in a big round hole because the shapes agree, but you can fit a small square peg in a big round hole; that is, the mismatched shapes permit only a small square peg to fit in. Matched pairs resemble the large round peg fitting the large round hole: in a matched pair, a serious unmeasured bias without a treatment effect could push the person with the higher response into the treated group. With matched triples, one treated subject and two controls, there can be a mismatch in shape forcing the worst possible bias to be smaller. In a matched triple with three different responses, a serious bias does not quite fit properly: it is unsure whether to push the individual with the middle response toward treatment or toward control. This does not happen with binary responses because there is no middle response, and a serious bias can push everyone with a 1 response toward treatment and everyone with a 0 response toward control. In other words, three very different responses in a matched triple tell you that the most serious possible bias does not fit the observed situation—some of the variation in responses, r_{Ci}, cannot be closely tracking the treatment, Z_i, because the treatment and outcome distributions fit together imperfectly.[33]

The increase in design sensitivity, $\tilde{\Gamma}$, from using $K = 3$ controls rather than pairs is not large, but neither is it trivially small. Design sensitivity is a limit as the sample size increases; however, in finite samples, the increase in both the design sensitivity, $\tilde{\Gamma}$, and sample size from increasing K can have a material effect on the power of a sensitivity analysis.

✱   Matching to a Variable Number of Controls

Instead of matching each treated individual to the same number, K, of controls, the number of controls may vary. In the United States in 2016, cigarette smoking is uncommon among people with more education and more income, but is not uncommon among people with less education and less income. As a result, there are many more controls available for a smoker with more education and income, and fewer available for a smoker with less education and income. In this case, one might match a smoker with more education and income to, say, five controls with more education and income, but one might match a smoker with less education and income to, say, two controls with less education and income. If the number of controls available changes with the observed covariates, x_i, then the number of matched controls may reasonably vary with x_i.

How many controls do we expect to be available at a given value of the observed covariate, x? Frank Yoon proposed a simple answer called the "entire number."[34] The entire number is defined to be $(1 - \lambda_x) / \lambda_x$, where λ_x is the propensity score. If, at a given x, the propensity score is $\lambda_x = 1 / 3$, then at this x we expect one-third of people to receive treatment and two-thirds to receive control, so for every treated individual we expect to see two controls, or an entire number of $(1 - \lambda_x) / \lambda_x = (1 - 1 / 3) / (1 / 3) = 2$. One might reasonably view the entire number as a ceiling on the number of controls to be expected at a given x; however, close matches may require that fewer than $(1 - \lambda_x) / \lambda_x$ controls be used.

Suppose that one has 100 treated individuals, 600 potential controls, and one has committed to using 300 of the 600 controls in the matched sample. In this case, one could match 1-to-3 in all cases. As an alternative, one could use 300 controls but permit matched sets with a blend of sizes 1-to-2, 1-to-3, and 1-to-4. What are the advantages of these two options? The steady 1-to-3 match has a small advantage in producing a stable estimate, with a smaller variance. Generally, the closest match with variable numbers of controls will be closer in terms of covariate distances and will be able to remove more bias from observed covariates.[35] Writing up results for a 1-to-3 match is easier than for a variable match. In a 1-to-3 match, every control counts the same as every other control, so one computes a mean or a boxplot for controls in the usual way. In a variable match, each control in a 1-to-2 matched set counts

for twice as much as each control in a 1-to-4 matched set, with the consequence that even descriptive statistics and boxplots describing the control group require some care and effort.[36]

* Full Matching

A step beyond matching with a variable number of controls is "full matching."[37] In full matching, each matched set consists of either (i) one treated individual and one or more controls or (ii) several treated individuals and one control. Full matching is useful when, at some value of the observed covariates, x, there are more controls than treated individuals, while at other values of x there are more treated individuals than controls. Saying the same thing in a different way: full matching is useful when Yoon's entire number, $(1 - \lambda_x) / \lambda_x$, can be either greater than or less than 1.

There is a specific sense in which full matching is the optimal form for an observational study. To avoid technical details, the description that follows is slightly informal.[38] Suppose that you had no particular interest in matching, and you simply wanted to compare treated and control individuals in strata, like those in Table 5.1. Suppose, as is typically true, that there are many observed covariates in x, including some continuous covariates such as weight, so no two people are absolutely identical in terms of x; therefore, the covariate distance between any two people is a positive number (recall again Table 5.6). Of course, we do not want strata in which everyone received treatment or in which everyone received control because there is nothing to compare in such a stratum. So we want people in the same stratum to be as similar as possible in terms of x, but we do not want to get carried away and create many strata that do not permit comparisons. Consider the following problem: find strata to minimize the average distance between treated and control individuals in the same stratum, with the requirement that every stratum has someone to compare—that is, every stratum has at least one treated person and at least one control. The solution is a full matching. In other words, if you start off wanting good strata, not a matched comparison, and you optimize the strata, then you end up with a full matching. The proof of this conclusion is a little technical in its details, but the idea of the argument is easy to understand. One demonstrates that if

you had a stratum with two or more treated individuals and two or more controls, then you could reduce the average distance within strata by splitting that one stratum into two strata. So no matter where you start, you are pushed to split strata until you can split no more, and the result is a full matching.

In principle, full matching permits every treated individual and every control to be matched. However, even with full matching, it is often wise to use some but not all controls, as illustrated in the following example.

An interesting study that used full matching concerned the effect on a student of being retained in the first grade of elementary school.[39] Is it of benefit to the student to retake grade one if the student is lagging behind? Or would the student be better off continuing on to the second grade? Wei Wu, Stephen G. West, and Jan N. Hughes began with 784 students in first grade in three schools in Texas, of whom 124 were retained in the first grade. The 124 retained students were matched for 72 observed covariates, x_i, to 251 promoted students. The full match permitted one retained student to be matched to between 1 and 5 promoted students, or one promoted student to be matched to between 1 and 5 retained students. As one might expect, before first grade ended, many of the $660 = 784 - 124$ students who were not retained were very different from the 124 retained students—many of these 660 students were at no risk of retention—so it makes sense to focus on 251 promoted students who more closely resemble the retained students. The students with low probabilities of retention based on observed covariates, x—that is, students with low propensity scores λ_x—comprise a small part of the matched retained and promoted groups, but they comprise a large part of the $409 = 660 - 251$ excluded promoted students. The use of full matching takes the adjustment a step farther: a retained student with a very high propensity to be retained, λ_x, is likely to have available fewer similar promoted students to serve as controls than a student with a lower propensity to be retained because the first student has a smaller entire number, $(1 - \lambda_x)/\lambda_x$. The flexible matching ratio permits the number of controls used in matching to adjust to the controls that are available. The study judged that the retained students benefited over several years from being retained in first grade, in terms of teacher-rated hyperactivity and behavioral engagement; however, the study also expressed some concerns about whether benefits outweigh other considerations in the longer term.

Jennifer Hill, Jane Waldfogel, Jeanne Brooks-Gunn, and Wen-Jui Han used full matching and other methods in a study of the effects maternal employment on a child's early development, finding some negative effects on a child's cognitive outcomes for mothers who worked full-time in the child's first year.[40] Phuong Nguyen-Hoang used full matching and other methods in a study of voter referendums on public school expenditures in small towns in New York State, concluding that referendums reduced spending on public education mostly by increasing class size but without reducing the cost of administrative overhead.[41]

✳ Risk Set Matching

In the typical randomized experiment, each person is assigned to treatment or control upon entry into the study. In contrast, in more than a few observational studies, everyone enters the study as untreated, and periodically someone who has been in the study for a while is given the treatment. In this context, it is a substantial mistake to compare everyone who received treatment to everyone who received control as if they were treated and control groups.

The best illustration of the problem comes from early studies of the effects of heart transplantation. A person who needed a new heart was immediately entered into the study but had to wait until a suitable heart became available. The early studies compared the duration of survival from entry into the study for people who were transplanted to those who were not transplanted. Can you spot the problem with this?

In 1972, Mitchell Gail wrote an influential critique of the early studies of heart transplantation.[42] He asked us to consider the fate of a patient who dies before a heart becomes available. By definition, that patient is a "control," a patient who did not have a heart transplant. More or less by definition, that patient did not survive very long. In contrast, every transplanted patient survived for a while, at least long enough to receive a heart. Even if no transplants had been performed—even if a patient was simply labeled "transplanted" when a heart became available, with the heart discarded and no operation performed—this way of constructing transplant and control groups would lead us to expect that the transplant group will survive longer in aggregate. If Fisher's hypothesis of no effect of transplant were true, then

this method of constructing treated and control groups would produce longer survival in the transplanted group.

Sally, patient $i = 1$, entered the study on the first of the month, and a heart became available two months later, whereupon Sally received a transplant. She survived for an additional 12 months, so her survival under transplant was $r_{T1} = 2 + 12 = 14$ months from her entry into the study. What would her survival, r_{C1}, have been had she not been transplanted? We do not know, and that is how it should be. In contrast, poor Harry, patient $i = 2$, entered the study on the first of the month and died a month later, before a heart became available, so Harry survived $r_{C2} = 1$ month from entry into the study. At the end of the second month, a heart arrived that would have been suitable for Harry. How long would Harry have survived had he received a heart transplant? Assuming that transplanting Harry at two months would not have resurrected him, we know that his survival would have been $r_{T2} = 1$ month from entry into the study. In truth, there really is no causal effect—no counterfactual—for Harry, because Harry never had a chance to be transplanted. A counterfactual for Harry would have to postulate that he lived longer than he did and received a heart, but if you could make patients live longer by postulating, then surgeons would not be needed. The early studies of heart transplants erred by putting people like Harry in the control group—in truth, they never had a chance.

How would one conduct a randomized experiment if people had to wait for a while before receiving treatment? Gail suggested an experiment similar to the following randomized experiment. When a heart becomes available, two similar people who are appropriate candidates for that heart are paired, and one is picked at random to receive it. At that moment, the clock starts for this pair, and survival is measured from the moment of randomization. This experiment is an equitable comparison—transplanted patients are expected to survive longer only if transplantation prolongs survival.

The paired randomized experiment just described comes in two versions. In version one, the patient who does not receive a heart at randomization never receives one, is never transplanted. This first version was discussed by Gail. In the second version, the patient who does not receive a heart at randomization might or might not receive one later, so the experiment estimates the effect of transplantation now versus delaying transplantation, waiting and seeing, and possibly transplanting later. The second version is, in certain respects, analogous to some observational studies.

Risk set matching is a design for an observational study that is analogous to the second version of the paired randomized experiment.[43] At the moment that one person receives treatment, that person is paired with someone else who has not yet been treated whose observed covariates were similar up to that moment. Importantly, the matching controls for the past, not for the future. That is, risk set matching makes people similar before treatment, but it does not remove part of the treatment effect by making them similar after treatment.[44] The untreated person in this pair may never be treated or may be treated the very next day. Because the matching respects the temporal sequence of events, it avoids the problems that occurred in the early heart transplant studies. As always, matching for observed covariates may fail to control some unobserved covariate—risk set matching does not eliminate that basic problem.

∗ Matching within and between Institutions That Provide Treatment

Often people receive treatments at institutions, such as hospitals, schools, prisons, or corporations. People find their way to a treatment by finding themselves at a particular institution and being selected for that treatment within that institution. These two steps may each introduce biases into treatment comparisons, but they will often be different biases. Perhaps we end up at a particular hospital or school by virtue of where we live, and we receive a particular treatment at that institution because of the way that institution assigns treatments to people who come within its orbit.

For instance, José Zubizarreta and colleagues compared the effects of general and regional anesthesia for knee replacement surgery.[45] During regional anesthesia, pain is blocked, but the patient remains conscious, perhaps sedated. During general anesthesia, the patient is not conscious. Both forms of anesthesia are feasible and common for knee surgery. The study looked at patients from 47 hospitals. At some hospitals most patients received general anesthesia, at other hospitals most patients received regional anesthesia, and at still other hospitals both forms of anesthesia were commonly used. Is one form of anesthesia better than the other? The study looked at various outcomes in various combinations, including mortality, deep-vein thrombosis, and readmission to the hospital within 30 days.

The study created two matched comparisons. The two comparisons did not overlap—no patient was used twice—so the results of the two comparisons are statistically independent. In the first comparison, if a hospital contributes 22 patients to the general anesthesia group, then it also contributes 22 patients to the regional anesthesia group; that is, the 47 hospitals were perfectly balanced. If one hospital is consistently better than another, then that does not bias this first comparison because this hospital appears in the general and regional groups with equal frequency. For brevity, the first matched comparison is called the "finely balanced" match, because the 47 hospitals are perfectly balanced. The worry about this first comparison is that someone or some process inside the hospital gave general anesthesia to Harry and regional anesthesia to Sally, and we as investigators do not know how or why this decision was made.

In the second matched comparison, a hospital contributed patients to either the general anesthesia group or the regional anesthesia group, but not to both groups. In the second matched comparison, the "usual practice comparison," hospitals contributed patients only to the group that represented the more common practice in that hospital. A hospital that typically used general anesthesia contributed patients to the general anesthesia group, while a hospital that typically used regional anesthesia contributed patients to the regional anesthesia group. In the usual practice match, hospitals are seen doing what they usually do. The worry is that the second comparison is affected by other differences among hospitals that may be associated with their choice of anesthesia.

In effect, there are two unrelated studies, each free of one problem that affects the other. It would be reassuring to reach similar conclusions from two independent studies with different weaknesses, and it would be a source of concern if they pointed to different conclusions. Because the two studies do not overlap and are statistically independent, their results can be combined using methods for combining the results of independent studies. In effect, the matching has created two evidence factors, as discussed in Chapter 7.

Table 11.3 describes the two comparisons. The observed covariates x contained 44 variables plus the hospital identifier with 47 levels. Table 11.3 describes 7 of the 44 variables and 6 of the 47 hospitals. The finely balanced match contained 1,354 matched pairs, and the usual practice match contained 944 matched pairs. Each match is balanced on its own, but there is

no reason the two matches should be similar to each other. At the bottom of Table 11.3, it is seen that hospital 1 contributed 154 = 77 + 77 patients to the finely balanced match: 77 patients undergoing general anesthesia and 77 patients undergoing regional anesthesia. Hospital 1 also contributed 11 patients undergoing regional anesthesia to the usual practice match. In contrast, hospital 47 contributed 44 = 22 + 22 patients to the finely balanced match: 22 patients undergoing general anesthesia and 22 undergoing regional anesthesia. Hospital 47 also contributed 97 patients undergoing general anesthesia to the usual practice match.

Table 11.3.   Matching within and between institutions that provide treatment

Anesthesia type	Finely balanced match (1,354 pairs)		Usual practice match (944 pairs)	
	General	Regional	General	Regional
Sample size	1,354	1,354	944	944
Age (years, mean)	72.4	72.5	72.6	72.4
BMI < 18 (%)	0.3	0.3	0.0	0.0
BMI > 30 (%)	52.4	52.4	54.1	54.1
Systolic BP (mean)	143.0	143.0	142.5	142.9
APACHE II score (mean)	22.4	22.6	22.3	22.4
Congestive heart failure (%)	6.5	5.8	4.1	5.2
Past MI (%)	4.3	3.9	3.9	4.1
Plus 37 additional covariates x				
Hospital, 1 to 47	**General**	**Regional**	**General**	**Regional**
1	77	77	0	11
2	18	18	0	66
3	48	48	0	10
4	35	35	168	0
⋮	⋮		⋮	
46	73	73	0	4
47	22	22	97	0

In the finely balanced match, each hospital contributed the same number of patients to the general anesthesia and regional anesthesia groups. In the usual practice match, a hospital contributed patients to only one group, representing the more common practice in that hospital.

APACHE II = Acute Physiology and Chronic Health Evaluation; BMI = body mass index; BP = blood pressure; MI = myocardial infarction.

The match in Table 11.3 is constructed by an optimization algorithm that builds the largest possible finely balanced match, then optimally constructs the usual practice match from patients not yet matched.[46]

✱ Template Matching

Template matching was proposed as a tool for auditing the health outcomes, costs, and value provided by many hospitals. Silber and colleagues used template matching to compare 217 hospitals in New York, Illinois, and Texas in terms of cost and quality in gynecologic, urologic, orthopedic, and general surgery.[47] Viewed abstractly, template matching is a tool for comparing many treatments to one another.

In their study, a template of 300 surgical patients was constructed by sampling patients from the pooled patient population of the 217 hospitals. Then patients from each of the 217 hospitals were matched to the template. This produced an array with 300 rows and 217 columns, each row corresponding with a different patient in the template, each column being a different hospital. The array contains $65,100 = 300 \times 217$ patients. Within a row, the patients are similar by virtue of being matched to the same patient in the template. In this way, one compares 217 hospitals in terms of their surgical performance on 300 similar patients.

When evaluating a single hospital, that hospital had 300 patients, each one matched to 216 controls from other hospitals, a comparison of 300 patients to $300 \times 216 = 64,800$ controls. An earlier section discussed some of the strengths and limitations of matching with multiple controls.

In terms of measured covariates, x, the patients in the 217 hospitals or columns were far more similar than if patients had been randomly assigned to hospitals. Despite this, there were substantial differences among the 217 hospitals in outcomes, that is, in mortality rates, costs, and readmissions to the hospital within 30 days.[48] Our daily experience of purchasing goods and services in competitive markets leads us to expect that better quality costs more. This expectation is not confirmed when comparing hospitals. A sensitivity analysis suggested that the pattern cannot be explained by small unmeasured biases in the assignment of patients to hospitals.[49] Dying in the hospital is expensive, so preventing excess deaths is a plausible strategy for controlling excess costs.

Taking Stock

Matching creates a simple comparison of treated and control groups that looked similar before treatment in terms of measured covariates, x. Modern techniques are a kit of tools to achieve specific objectives. Some tools, such as fine balance, improve adjustments for observed covariates, x. Other tools—such as pairing aimed at reducing heterogeneity of outcomes—attempt to reduce sensitivity to unmeasured covariates, u. Still other tools, such as risk set matching, address problems that often arise when the investigator does not control the timing of treatment.

Biases from General Dispositions

What Are Generic Unobserved Biases?

Rashness, Caution, Wisdom, Impulsiveness, Integrity, and Other Generic Biases

In everyday conversation, we often attribute a particular choice made by a person to a disposition of that person to make choices in a particular way. Harry does not wear seatbelts, texts when driving, often tailgates, and does not wear his helmet when cycling because Harry is a reckless person. Sally wears seatbelts, never texts when driving, never tailgates and wears a helmet when cycling because she is a cautious person. Although we may err in attributing a disposition to a person, presumably we do not always err, and talk of dispositions does some work for us. Who is more likely, Harry or Sally, to cycle though a red traffic signal without stopping? Presumably, from the information given, we would guess it is Harry, but we could be mistaken here. In addition to erring about an additional behavior, we may err in characterizing the disposition, perhaps defining it too broadly. As it turns out, Harry keeps his money in the bank while Sally buys stock options; that is, Harry is reckless about personal safety but very cautious with money, and Sally is the opposite. The lines between rashness, impulsiveness, thoughtlessness,

and stupidity can be difficult to draw, yet nonetheless it may be a reasonable bet that Harry will cycle through a red traffic signal while Sally comes to a full stop.

Does a person take a single precaution? Perhaps some people do, but many people exhibit behaviors that fit together in consistent patterns. Dispositions may have intelligible origins. Harry got his driver's license in 1985, and Sally got hers in 1986. Beginning in 1986, to get a driver's license you had to watch an intense film depicting a horrible, fatal accident; ever since then the possibility of such an accident has been salient for Sally but not for Harry. In psychology, an extensive literature is devoted to testing whether a particular pattern of behavior is correctly attributed to a particular unmeasured, organizing disposition.[1]

That people behave in consistent ways is a problem in observational studies. Their behavior is better organized than you can see in your data, better organized than in any data anyone will ever see. If some people take many precautions and others take few, isolating the effects of a single precaution is not going to be easy. You would like to study the effects on injury severity of wearing a helmet while cycling, so you end up comparing Harry's injuries to Sally's. Your data record that Harry and Sally each fell from their bicycles and Harry's injuries were more severe, but the data do not record that Sally hit a bump at a slow speed and landed in the grass, while Harry sped through a red signal, dodged a car, lost control, and slammed into a lamppost. True, Sally but not Harry was wearing a helmet, but the speeds, grass, and lamppost matter also, and they are not in your data.

That people behave in consistent ways is a problem in observational studies. The more consistent people are, the greater the problem. Perhaps we should take a special interest in people who are less consistent, less predictable than Harry and Sally. Harvesting inconsistent elements from consistent patterns of behavior is the ambition of this chapter.[2]

A generic unobserved bias is an unmeasured covariate u that promotes several treatments at the same time in a parallel way. If u is a disposition to behave in ways that promote personal safety, then higher values of u may promote wearing a helmet while cycling, stopping at red signals, wearing a seatbelt, and not texting while driving, and so on. A person like Sally who has a high value of u may do many things that promote personal safety, while a person like Harry with a low value of u may do many things that promote risk. If you compare Harry and Sally in a study of the effects of bicycle

helmets, you have made a very biased comparison. The data you have or can ever obtain will always omit many of the risks Harry took and Sally avoided.

Are There Strategies That Can Remove Unmeasured Generic Biases?

The surprising thing about generic unobserved biases is that they can be broken up, sometimes reduced, and sometimes removed entirely. By definition, generic unobserved biases affect many behaviors, and although you have not observed most of these behaviors, you may have observed a few. How can this fact be put to use?

Harry and Sally are unhelpful because their behavior is highly consistent. Enter David and Debbie, who are less consistent. If randomized treatment assignment is the goal, "less consistent" is a small step in the right direction. David never texts while driving—that would be totally irresponsible—but he skips the helmet while cycling because he likes the wind in his hair. Debbie never skips the helmet while cycling—that would be totally irresponsible—but she does text while driving because she is good at multitasking. In terms of concern with personal safety, David and Debbie both fall between—are more alike—than Harry and Sally. It is harder to guess whether David or Debbie would cycle through a red traffic signal, and easier to guess that Harry would while Sally would not. Everything Harry does says he is indifferent to personal safety, and everything Sally does says she is concerned about it. For David and Debbie, personal safety is sometimes a concern, but it does not override all other preferences. Expressed in terms of dispositions, we cannot see the disposition u for personal safety for any of these four people, but looking at the behavior that we can see, we guess that u is low for Harry, high for Sally, and somewhere in between for David and Debbie. Is this useful in studying the effects on injury severity of wearing a helmet while cycling?

It is indeed, because two key aspects are both present. First, in an observational study it is a good thing to be unable to guess whether David or Debbie is more likely to cycle through a red traffic signal. Comparing David with Debbie may be biased in ways we cannot see, but at least the unmeasured biases are less obvious than when comparing Harry and Sally—everything we see says the latter is a biased comparison. Second,

not texting while driving and wearing a helmet while cycling are both indicators of a disposition toward personal safety, but only the helmet can cause greater safety while cycling. In comparing the cycling injuries of David and Debbie we are comparing two people who have a middling concern for personal safety, something that may affect relevant behaviors we have not observed, such as cycling speed and stopping at red signals. At the same time, in comparing the cycling injuries of David and Debbie, we are comparing one person who wore a helmet and one who did not.

David and Debbie form a better treatment–control pair than Harry and Sally, yet we paired David and Debbie precisely because they were different, not because they were the same. David never texts while driving, but Debbie does. An unthinking, mechanical investigator would have paired David with a helmeted control who never texts while driving, perhaps Sally; then David would resemble his control in terms of texting. We know that if you text while driving it will not cause you to be distracted while cycling. We paired David and Debbie because we took their visible behavior to indicate a middling disposition u to personal safety—one cycling without a helmet and the other texting while driving. We expected that middling disposition u to personal safety to manifest itself in ways we did not measure: cycling speed and stopping at red traffic signals. We paired David and Debbie despite their visible differences in safety behavior because we took their patterns of safety behavior to suggest that they were similar in terms of a generic disposition toward personal safety. David and Debbie are twice different—helmet and texting—but we took that as a sign that, beneath the surface, they are much more similar to each other than Harry and Sally.

Differential Effects of Two Treatments

If there are two treatments, the differential effect of the two treatments is the effect of receiving one treatment in lieu of the other. For example, one treatment is texting while driving. Another treatment is wearing a helmet while cycling.

In principle, the differential effect of two treatments could be very different from the effect of either of them. Bayer aspirin might cure your headache, and Walmart aspirin might cure your headache, but the differential effect of Bayer aspirin in lieu of Walmart aspirin might be zero, because they

are equally effective at curing headaches. Keep that firmly in mind: the differential effect of two treatments need not equal the effect caused by either of them.

Despite this, differential effects are sometimes highly instructive. Texting while driving could cause your death while driving, but presumably it has no effect on injuries you sustain while cycling. In light of this, consider a differential effect on the severity of injury sustained while cycling. One treatment is wearing a helmet while cycling. The other treatment is not texting while driving. In general, the differential effect of two treatments is not the effect of either one, but is that true here? Presumably, texting while driving does not cause any effect on injury severity while cycling. If so, the differential effect on cycling injuries of wearing a helmet while cycling in lieu of texting while driving is just the effect of wearing a helmet while cycling.

Let us put together the pieces. We liked the comparison of David and Debbie because their behavior exhibited a middling concern for personal safety, unlike Harry and Sally. David does not text while driving, but he does not wear a helmet either. Debbie wears the helmet but texts while driving. In comparing David and Debbie we are estimating the differential effect of helmets and not texting, but we strongly suspect that differential effect is just the effect of helmets, so far as bicycle injuries are concerned. If all of this were true, it would be what we wanted: an estimate of the effects of bicycle helmets with reduced bias from a general disposition toward personal safety.

Behind the stories about Harry, Sally, David, and Debbie is a piece of mathematics.

Rasch Behavior

A Numerical Illustration of the Rasch Model for Treatment Assignments

Georg Rasch proposed an influential model for binary responses, correct or incorrect, to a psychological test or measure with several or many questions.[3] The model involves an unobserved variable u_i for person i and a parameter for each question. A person with a higher u_i is more likely to correctly respond

to every question, and some questions are harder than others, as reflected in the parameters that describe the questions. The model has a simple mathematical form in which the relative difficulty of two questions is the same for all people.[4] Sally has a higher u than Harry, but both find question 1 easier than question 2: Sally is twice as likely to get question 1 right as to get question 2 right, and the same is true for Harry.

The Rasch model is the simplest model that exhibits a certain interesting behavior. For us, the binary responses are not responses to questions but rather behaviors exhibited, such as wearing a helmet while cycling, Z, or not texting while driving, Z^*, where a 1 indicates the safer behavior and a 0 indicates the less safe behavior. Table 12.1 is an invented illustration of distributions of the two treatments, (Z,Z^*), under a Rasch model, for two people, one person with a high u like Sally, at the top of Table 12.1, the other person with a low u like Harry at the bottom of Table 12.1.[5]

In Table 12.1, the cautious person with a u_i like Sally has a 0.9 probability of not texting while driving, a 0.75 chance of wearing a helmet while cycling, and a 0.675 chance of doing both. In contrast, the less cautious person with a u_i like Harry has a 0.75 probability of not texting while driving, a 0.50 chance of wearing a helmet while cycling, and a 0.375 chance of doing both. We think that the person like Sally is more likely to behave cautiously in other ways including ways we do not see, and the person like Harry is likely to behave less cautiously in other ways. For example, we guess that Sally is more likely to stop cycling at a red traffic signal and that Harry will ride through, but those events are not in our data.

Although we cannot see the disposition, u_i, we can eliminate it using the observable behavior we can see, namely (Z_i, Z_i^*). That is remarkable. Typically, we cannot remove bias from an unobserved covariate, u_i, but in a situation like Table 12.1 we can. Suppose we make the differential comparison, comparing someone with $(Z_i, Z_i^*) = (0,1)$ like David to someone with $(Z_i, Z_i^*) = (1,0)$ like Debbie. That is, we compare someone like David, who neither wears a helmet while cycling nor texts while driving, to someone like Debbie who does both. In comparing David and Debbie, we compare two people who take some risks and avoid others. Here is the key fact: in both the upper and low part of Table 12.1, the ratio of the probability of behaving like David to the probability of behaving like Debbie—the ratio of the probability that $(Z_i, Z_i^*) = (0,1)$ to the probability that $(Z_i, Z_i^*) = (1,0)$—is the same, either $0.225 / 0.075 = 3$ in the upper part of Table 12.1 or

Table 12.1. Numerical illustration of treatment assignment probabilities from the Rasch model

		Texting while driving, Z*		
High u *More concerned with personal safety*		*No texting,* *Z* = 1*	*Texting,* *Z* = 0*	*Total*
Helmet while cycling Z	Helmet, $Z = 1$	0.675	0.075	0.750
	No helmet, $Z = 0$	0.225	0.025	0.250
	Total	0.900	0.100	1.000
Low u *Less concerned with personal safety*		*No texting,* *Z* = 1*	*Texting,* *Z* = 0*	*Total*
Helmet while cycling Z	Helmet, $Z = 1$	0.375	0.125	0.500
	No helmet, $Z = 0$	0.375	0.125	0.500
	Total	0.750	0.250	1.000

$0.375 / 0.125 = 3$. Under the Rasch model, this is true for all values of u_i, not just the two values in Table 12.1: every person is 3 times as likely to behave like David rather than like Debbie. Even Harry and Sally, with their very different dispositions u_i, are both 3 times more likely to behave like David than like Debbie. If the Rasch model in Table 12.1 governed treatment assignments—a big if—but we focus on the subset of people with $Z_i + Z_i^* = 1$, then in this subset we have a randomized experiment with a 0.75 chance of neither wearing a helmet nor texting and a 0.25 chance of doing both; that is, the bias from u_i has vanished.[6]

There is another remarkable fact. Suppose Z_i^{**} indicates some other behavior, such as stopping the bicycle at a red traffic signal, so the three behaviors (Z_i, Z_i^*, Z_i^{**}) follow the Rasch model. Intuitively, without the Rasch model, we thought Harry was likely to run red lights and Sally was unlikely to do so, but it was hard to guess whether David or Debbie would be more likely to do so. Actually, that is exactly correct under the Rasch model: given that we are talking about someone who behaved like David or Debbie with $Z_i + Z_i^* = 1$, the patterns $(Z_i, Z_i^*) = (0,1)$ for David or $(Z_i, Z_i^*) = (1,0)$ for Debbie contain no information about Z_i^{**}, so they are equally likely to run a red light.[7] That is, the differential comparison tends to balance other behaviors governed by the same disposition u_i under the Rasch model whether these behaviors are observed or not.

Essentially, what we have found is that the stories about Harry, Sally, David, and Debbie are correct as mathematics under the Rasch model. Study bicycle helmets in isolation, and you compare Harry and David, who do not wear helmets, with Sally and Debbie who do, but that comparison is very biased because Harry is doing many risky things because of his low u_i and Sally is doing very few risky things because of her high u_i. In contrast, if you look at the differential comparison, comparing David with Debbie, you compare two people with middling concern for personal safety, and the bias from u_i is gone. True, when you discover that David's injuries when cycling are more severe than Debbie's injuries when cycling, you do have to decide whether to attribute the effect to Debbie's helmet or David's not texting when driving, but that is not a tough call.

Examples

Nonsteroidal Anti-inflammatory Drugs and Alzheimer's Disease

There is a theory, perhaps correct, that long-term use of nonsteroidal anti-inflammatory drugs (NSAIDs) such as ibuprofen (e.g., brand name Advil) reduces the risk of Alzheimer's disease.[8] Studies of this topic face a variety of challenges. Long-term, accurate records of the use of NSAIDs are difficult to obtain, in part because many NSAIDs are sold over the counter, which does not produce an electronic record of either prescriptions or purchases. Even people without Alzheimer's disease may have difficulty remembering how many ibuprofen tablets they have taken over the previous decade. Alzheimer's disease develops gradually and may affect cognition long before it is clinically diagnosed, perhaps affecting use of pain relievers. Specifically, in 't Veld and colleagues write, "Pain perception and expression may be different in those becoming cognitively impaired or in demented subjects. If either pain perception or expression is impaired in those developing Alzheimer's disease, this impairment may lead to a lesser use of NSAIDs and an ostensible protective effect of NSAIDs."[9] In other words, perhaps a person in the early, undiagnosed stages of Alzheimer's disease is less aware of pain or is less inclined or able to act effectively in response to pain, so that person may

consume fewer pain relief medications than someone without Alzheimer's disease. This could create a spurious association between greater use of medication and reduced incidence of clinically diagnosed disease.

The potential bias just mentioned is generic: if it is real, it should depress use of pain relievers of all kinds. Not all pain relievers are NSAIDs. For instance, acetaminophen (e.g., brand name Tylenol) is a pain reliever that is not an NSAID. Perhaps reduced awareness of pain or reduced ability to act to relieve pain reduces the use of pain relievers in people with undiagnosed Alzheimer's disease, but it seems less likely to induce a switch from ibuprofen to acetaminophen. Therefore, some studies have tried to break up the generic bias that should depress use of most pain relievers by looking at the differential effect of NSAID pain relievers and non-NSAID pain relievers. For instance, Peter Zandi and colleagues did this, finding a negative association between NSAIDs and Alzheimer's disease and no association between non-NSAID pain relievers and Alzhiemer's.[10]

Once again, the thematic point is that certain generic biases promote or depress multiple treatments, thereby creating biases when trying to estimate the effects of any one of these treatments. Differential effects contrast the effect of two treatments promoted by the same bias, for instance, NSAID and non-NSAID pain relievers. Certain differential effects—effects that contrast two of these multiple treatments—may be immune to the generic bias. Therefore, in favorable circumstances, a differential effect between two carefully selected treatments may be an aid in ruling out a particular generic bias.[11]

Seatbelts in Car Crashes

A difficulty in studying the effects of seatbelts in motor vehicle crashes is that people who wear seatbelts may drive more cautiously than people who do not wear them. If cautious drivers are more likely to wear seatbelts, to be sober when they drive, and to drive at a safe speed with ample distance from the car ahead, while drivers who are not cautious are less likely to wear seatbelts, less likely to be sober, and more likely to drive faster and closer—if wearing or not wearing seatbelts is part of a larger pattern of driving behavior—then people who wear seatbelts may be involved in less severe crashes. In motor vehicle crashes that involve at least one fatality, the U.S. Fatal Accident Reporting System (FARS) records use of seatbelts and injury severity, but it

contains little information about speeds, distances between cars, and so on, so it contains less information than one might like about the forces involved in the crash.[12] In any physical event, forces matter. How can the effects caused by seatbelts be isolated from the effects of a generally cautious style of driving?

In a clever study, Lawrence Evans compared drivers and passengers in the same car in the same crash, one belted and the other unbelted.[13] Because driver and passenger were in the same car in the same crash, the speed was the same, the distance from the car ahead was the same, the driver's reaction time was the same, and many other aspects of the crash were similar if not the same. It is unusual for driver and passenger to be differently belted, so the comparison focuses on a minority of crashes. In principle, an unbelted individual may be thrown about, causing injury to a belted individual, so this comparison does not estimate the absolute effect of seatbelts; rather, it looks at the differential effect in the presence of someone else who is not belted.[14] Nonetheless, Evans's strategy goes a long way toward neutralizing biases from a pattern of cautious driving.

Figure 12.1 follows Evans's approach with more recent data from the 2010–2011 FARS database in which either a driver or a right-front passenger wore a lap-shoulder belt while the other was unbelted.[15] In FARS, injuries are scored 0 to 4, with 0 being no injury and 4 being death. Figure 12.1 displays driver-minus-passenger differences in injury scores, so positive values indicate the driver was more severely injured, and negative values indicate the passenger was more severely injured. Notably in Figure 12.1, the difference in injury scores is lopsidedly negative when the driver is belted and the passenger in not, but the difference is lopsidedly positive when the driver is unbelted and the passenger is belted. Regardless of seat position, the unbelted individual tended to suffer more severe injuries.

Evans's comparison eliminates many unmeasured biases. However, within the same car in the same crash, someone could claim that the type of person who tends not to wear safety belts is the type of person who is prone to injury, that frail individuals do not wear safety belts. How sensitive is the differential comparison in Figure 12.1 to unmeasured biases that operate within the same car in the same crash? Using the techniques in Chapter 9, a substantial bias of $\Gamma = 5.1$ could not produce either panel of Figure 12.1 in the absence of an effect of safety belts on injuries.[16] In parallel with the final columns of Table 9.1, $\Gamma = 5.1$ corresponds with an unobserved covariate

A belted driver with an unbelted passenger, n = 1,412

An unbelted driver with a belted passenger, n = 1,198
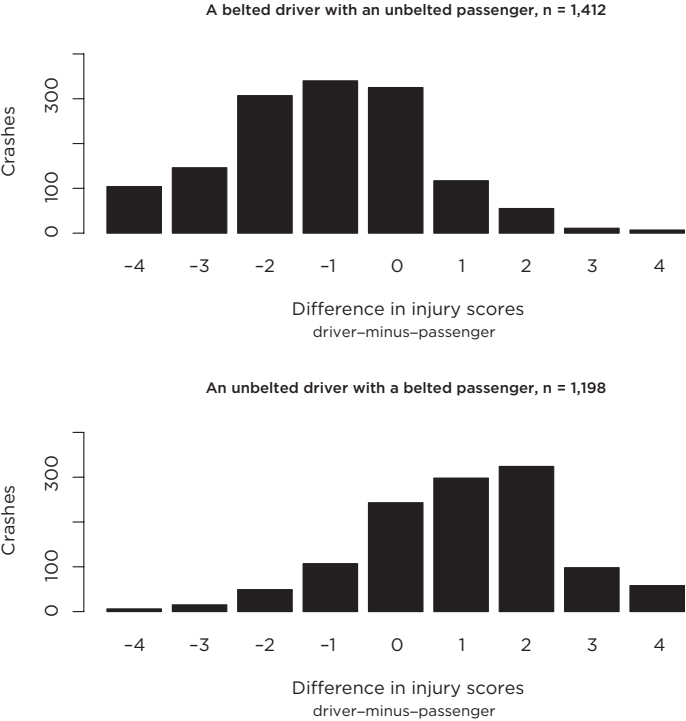
Figure 12.1.   Driver-minus-passenger difference in injury scores in crashes from the 2010–2011 U.S. Fatal Accident Reporting System in which the driver and front-right passenger were differently belted. Injury scores range from 0 = none to 4 = death; so (i) a driver-minus passenger difference of 4 means the driver died and the passenger was uninjured, (ii) a difference of −4 means the driver was uninjured and the passenger died, and (iii) a difference of 0 means the same injury for driver and passenger.

associated with a tenfold increase in the odds of not wearing a safety belt and a tenfold increase in the odds of more severe injuries than the traveling companion in the front seat.

Lead and Cadmium in the Blood of Cigarette Smokers

Does cigarette smoking cause an increase in the levels of lead and cadmium in a smoker's blood? Table 12.2 describes 518 matched pairs of a daily smoker and a nonsmoker from the 2009–2010 U.S. National Health and Nutrition

Examination Survey (NHANES).[17] In this comparison, nonsmokers smoked fewer than 100 cigarettes in their lives and none in the previous 30 days. Daily smokers smoked at least 10 cigarettes per day on every day of the previous 30 days. So these are real smokers and real nonsmokers, without equivocators.[18] The results in Chapter 10 suggested that a comparison is likely to be less sensitive to bias if it compares high doses with zero doses, excluding equivocators with modest doses.

Table 12.2 describes the balance on observed covariates, age, gender, education, race/ethnicity, and income measured as a ratio to the poverty level. The match exhibits the concept of fine balance from Chapter 11. For example, 258 of the 518 daily smokers are women, and 258 of the 518 nonsmokers are women, and the same pattern of perfect balance holds for education, race/ethnicity, and two income categories. Indeed, not seen in Table 12.2, there are six black women with a high school education and an income of more than twice the poverty level in both the smoking and nonsmoking groups, but they are not necessarily matched with each other; moreover, the same perfect balance exists in each of the $60 = 2 \times 2 \times 3 \times 5$

Table 12.2. Covariate balance in a matched comparison of 518 daily smokers and 518 never smokers. Tabulated values are counts, except as noted

	Daily smoker	Never smoker
Age (mean)	43.7	43.2
Female	258	258
Male	260	260
Gender Total	518	518
< ninth grade	43	43
≥ ninth grade, no diploma	119	119
High school or equivalent	170	170
Some college	152	152
BA degree or more	34	34
Education Total	518	518
Black	104	104
Hispanic	64	64
Other	350	350
Ethnicity Total	518	518
< 2 × poverty level	326	326
≥ 2 × poverty level	192	192
Income Total	518	518
Poverty ratio (mean)	2.0	1.9

categories formed from gender, income, race/ethnicity, and education. In addition, the pairs are as close as possible in terms of the covariates, including the numeric covariates of age and poverty ratio. Does that ensure that the smokers and nonsmokers are comparable?

Figure 12.2 depicts 518 smoker-minus-control pair differences in lead and cadmium. As in Chapter 8, Figure 12.2 reports the differences of base-2 logs, or $\log_2$, so that a difference of one means one-doubling, and a difference of two means two-doublings. Looking at the median or middle lines in the boxplots, it is seen that typically smokers have four times as much cadmium and twice as much lead in their blood as their matched nonsmoking controls, that is, two-doublings or one-doubling.

How sensitive are the comparisons in Figure 12.2 to failure to match for some unmeasured covariate? Where in Table 9.1 do these comparisons fall?
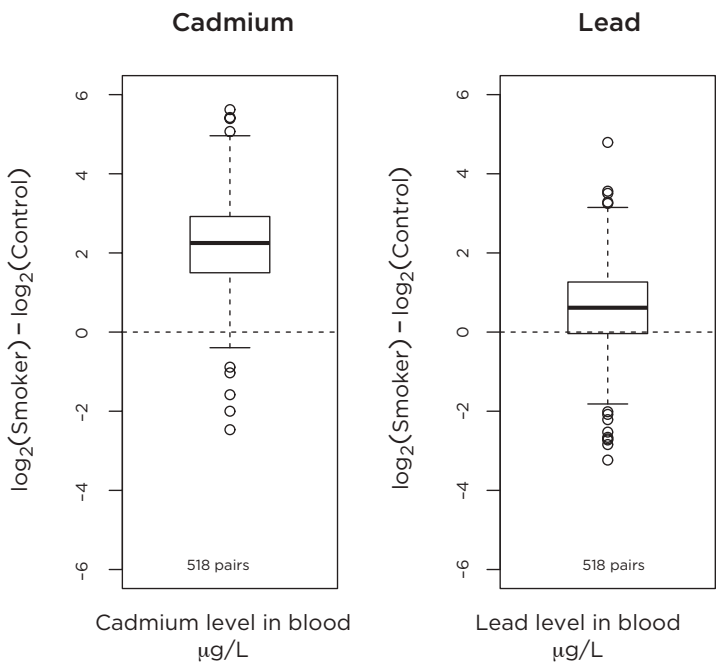
Figure 12.2. Matched pair comparison of levels of cadmium and lead in the blood of 518 daily smokers and 518 matched never-smokers. The values are 518 smoker-minus-control pair differences in the $\log_2$, so a difference of 1 unit is one doubling.

The answers are quite different for lead and for cadmium. For biased coin flips to produce the boxplot for lead in Figure 12.2, the coins would have to be quite biased, with the odds of a head being $\Gamma = 2.6$-to-1. In terms of the last two columns of Table 9.1, $\Gamma = 2.6$ corresponds with an unobserved covariate u_i that increases the odds of smoking by a factor of five and also increases the odds of a positive pair difference in lead by a factor of five. The comparison for cadmium is insensitive to enormous biases, becoming sensitive at $\Gamma = 64$ rather than $\Gamma = 2.6$.

Smoking cigarettes exposes an individual to a specific health hazard, but it also expresses a lack of concern about such hazards. Perhaps smokers do this in more than one way and are exposed to more than one hazard that nonsmokers typically avoid. Indeed, the NHANES survey provides strong evidence of this. The survey asks people whether they have ever tried cocaine, crack cocaine, heroin, or methamphetamine, which I will abbreviate as "hard drugs." Of the 1,036 people in Figure 12.2, 86% or 886 people answered that question. Of these, daily smokers were six times more likely than nonsmokers to have tried hard drugs, a substantial difference.[19] One could add this question about hard drugs to the list of observed covariates in Table 12.2 and match for it as well, but doing this would control for one more measured covariate, not for an unmeasured general disposition u_i to indulge hazardous habits. Matching for observed hazardous habits is never quite enough to compensate for a general disposition to indulge such habits.[20] Perhaps this is why we pay so much attention to manifestations of a person's disposition and character: there is always another manifestation coming, more of the same, waiting just out of sight around the corner. To adequately compensate for a latent disposition to indulge hazardous habits, we must overcompensate for the hazardous habits we can see, and that is what differential comparisons do.

Table 12.3 returns to the NHANES data, creating a new matched differential comparison of 105 daily smokers who never tried hard drugs compared with 105 never smokers who had tried hard drugs. Like David and Debbie, and unlike Harry and Sally, the differential comparison in Table 12.3 concerns people having an equivocal concern with their health, people whose behavior is a bit inconsistent. Table 12.3, like Table 12.2, shows that the matching was reasonably effective at balancing the specific covariates listed in the tables. The comparison in Table 12.3 overcompensates for observed

Table 12.3. Covariate balance in a differential matched comparison of 105 daily smokers who never tried hard drugs and 105 never smokers who tried hard drugs (tabulated values are counts, except as noted)

	Daily smoker who never tried hard drugs	Never smoker who tried hard drugs
Age (years, mean)	43.4	43.1
Gender		
Female	41	41
Male	64	64
Total	105	105
Education		
< ninth grade	5	5
≥ ninth grade, no diploma	15	15
High school or equivalent	17	17
Some college	50	50
BA degree or more	18	18
Total	105	105
Ethnicity		
Black	23	23
Hispanic	17	17
Other	65	65
Total	105	105
Income		
< 2 × poverty level	44	44
≥ 2 × poverty level	61	61
Total	105	105
Poverty ratio (mean)	1.8	1.6

habits in the hope of adequately compensating for a general disposition to indulge hazardous habits—it compares people who are visibly different in terms of hard drugs, hoping that they are similar in terms of an unmeasured disposition u_i.

In what sense does the differential comparison compensate not for having tried hard drugs but for an underlying penchant for risky health behaviors? Table 12.4 illustrates one aspect of the answer. Table 12.4 is analogous to the fictional story about who among Harry, Sally, David, and Debbie would cycle through a red light—that is, the fictional story about another unmeasured behavior governed by the same disposition. So far, we have not looked at alcohol consumption by the daily smokers and nonsmokers—it was not controlled by matching, nor did it play a role in defining

Table 12.4. Comparison of an unmatched covariate, alcohol consumption, in the basic and differential comparisons (values are percentages except as noted)

	Basic comparison		Differential comparison	
Number of drinks	Daily smokers (%)	Never smokers (%)	Daily smokers who never tried hard drugs (%)	Never smokers who tried hard drugs (%)
< 12 all year	12	36	10	9
1–2 on drinking days	31	32	39	41
3–4 on drinking days	28	17	23	22
> 5 on drinking days	29	15	28	28
Total	100	100	100	100
Count	385	412	100	94

the differential comparison. For people who responded to questions about alcohol, Table 12.4 uses the NHANES data to describe alcohol consumption in the four groups, the basic comparison in Table 12.2 with 518 pairs, and the differential comparison in Table 12.3 with 105 pairs. In Table 12.4, the basic comparison is out of balance: smokers drink more alcohol than nonsmokers, with 29% of smokers having more than five drinks per day, compared with 15% of never smokers. If a model similar to the Rasch model connected risky behaviors to an unmeasured disposition to indulge such behaviors, then that model would predict that alcohol consumption would be out of balance in the basic comparison and better balanced in the differential comparison. Although we do not know whether the Rasch model is correct here, we do see this pattern in Table 12.4: in the differential comparison, drinking behavior is similar for smokers who never tried hard drugs and for nonsmokers who had tried hard drugs. The important point is that the differential comparison balanced alcohol consumption without using alcohol consumption; rather, it overcompensated for having tried hard drugs.

Figure 12.3 parallels Figure 12.2 but now for the 105 matched pairs in the differential comparison. Notably, Figure 12.3 looks quite a bit like Figure 12.2. So when we compare two groups of people who have each indulged one risky behavior, it is only the smokers who have elevated levels of cadmium and lead in their blood, not the nonsmokers who tried hard drugs. This is consistent with smoking being the cause of these elevated levels of
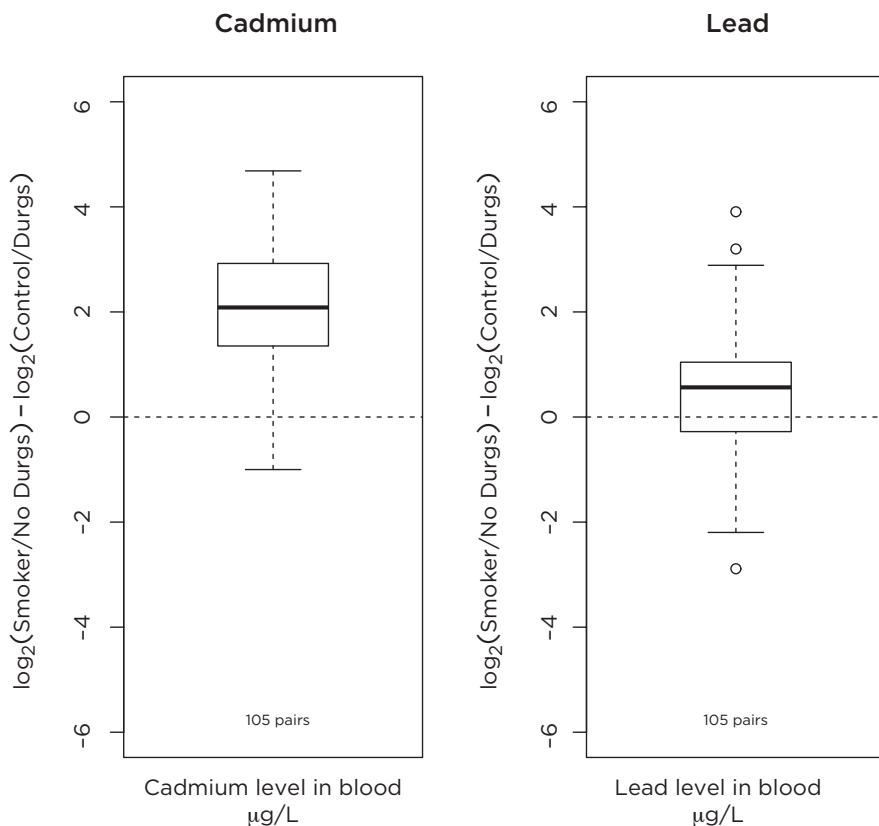
Cadmium   ## Lead

Figure 12.3. Differential matched pair comparison of levels of cadmium and lead in the blood of 105 daily smokers who never tried hard drugs and 105 matched never-smokers who tried hard drugs. The values are 105 smoker-minus-control pair differences in the $\log_2$, so a difference of 1 unit is one doubling.

toxins, as opposed to some other risky behavior also indulged by people with a penchant for behaviors that place their health at risk.

The comparisons in Figure 12.3 are immune to certain types of generic biases but are susceptible to biases that promote smoking without promoting the trying of hard drugs. For such differential biases, the comparison for lead in Figure 12.3 becomes sensitive to bias at $\Gamma = 1.8$, while the comparison for cadmium becomes sensitive at $\Gamma = 23$. At the far right in Table 9.1, to produce a bias of $\Gamma = 1.8$, an unobserved covariate u_i would have to both (i) increase the odds of smoking without having tried hard drugs by more than a

factor of three when compared with not smoking while trying hard drugs, and (ii) increase by more than a factor of three the odds of a positive difference in lead levels in a matched pair.

This example is intended to exemplify the claim that a basic comparison, like that in Figure 12.2, can be strengthened by adding a differential comparison, like that in Figure 12.3. This strengthening addresses a specific issue, that of isolating the effects of one behavior in the presence of a disposition that promotes several or many behaviors.

* Some Theory

The example in Table 12.1 referred to specific probabilities derived from a widely used model, the Rasch model. The Rasch model actually assumes much more structure than is needed for the conclusion we want, namely, that the differential comparison eliminates generic bias from an unobserved disposition u_i. The current section eliminates parts that are not needed for this conclusion. Then a sensitivity analysis addresses the possibility that the needed parts are absent.

Table 12.5 describes four possible treatment assignments for one person i with observed covariate x_i and unobserved covariate u_i. Here, there are two treatments, Z and Z^*, each at two levels, making four possible combinations of the two treatments, which person i receives with probabilities π_{11i}, π_{10i}, π_{01i}, and π_{00i}, where $1 = \pi_{11i} + \pi_{10i} + \pi_{01i} + \pi_{00i}$.[21]

Treatment assignment would be ignorable in Table 12.5 if two people, say, i and j, with the same observed covariates, $x_i = x_j$, had the same treatment assignment probabilities, $\pi_{11i} = \pi_{11j}$, $\pi_{10i} = \pi_{10j}$, $\pi_{01i} = \pi_{01j}$, and $\pi_{00i} = \pi_{00j}$. If treatment assignment were ignorable, we could compare any two treatments in Table 12.5 simply by matching for x.

Table 12.5.   Notation for one person in a differential comparison

		Texting while driving, Z^*	
		No texting, $Z^* = 1$	*Texting,* $Z^* = 0$
Person i with disposition u_i			
Helmet while cycling Z	Helmet, $Z = 1$	π_{11i}	π_{10i}
	No helmet, $Z = 0$	π_{01i}	π_{00i}

We are interested in the possibility that u_i does bias treatment assignments but promotes treatment Z and treatment Z^* in a similar way, and in this case we hope the differential comparison may be free of bias from u_i even though other comparisons are biased. By definition, there are only generic unobserved biases if two people, say, i and j, with the same observed covariates, $x_i = x_j$, have the same ratio $\pi_{10i}/\pi_{01i} = \pi_{10j}/\pi_{01j}$. A key point, illustrated in the very special case of the Rasch model in Table 12.1, is that treatment assignment probabilities, π_{11i}, π_{10i}, π_{01i}, and π_{00i}, may depend upon u_i, yet there may only be generic biases from u_i, in the sense that π_{10i}/π_{01i} does not depend upon u_i. When there are only generic biases, the differential comparison of $(Z=1, Z^*=0)$ and $(Z=0, Z^*=1)$ is free of unmeasured biases: the chance of $(Z=1, Z^*=0)$ rather than $(Z=0, Z^*=1)$ given that one of these has occurred does not depend upon u_i.[22]

The sensitivity analysis asks about violations of the assumption that there are only generic biases from u_i. That is, the sensitivity analysis allows π_{10i}/π_{01i} to depend upon u_i, but to a limited degree controlled by a parameter $\Gamma \geq 1$. Here, there is no restriction on how strongly π_{10i} or π_{01i} may vary with u_i—they may vary very strongly with u_i—but the sensitivity analysis restricts the relationship between π_{10i}/π_{01i} and u_i. Specifically, the sensitivity analysis assumes that two people, say, i and j, with the same observed covariates, $x_i = x_j$, may have ratios, π_{10i}/π_{01i} and π_{10j}/π_{01j}, that differ by a factor of $\Gamma \geq 1$ because individuals i and j differ in terms of the unobserved covariate, $u_i \neq u_j$. The differential comparison was free of bias if u_i only creates generic biases, but what if that assumption is wrong? We would like to study sensitivity to violations of that assumption. With a little attention to detail, we find that if you apply the sensitivity analysis in Chapter 9 to the differential comparison, then you have entirely eliminated generic biases and are now studying the sensitivity of that comparison to any biases that are not generic, biases that continue to affect π_{10i}/π_{01i}.

Isolation

What Is Isolation?

We have seen that the differential comparison of two treatments may not be biased by a disposition that promotes both treatments, even when the

comparison of either treatment to no treatment would be biased by that disposition. Being in pain may promote the use of pain relievers without biasing a comparison of ibuprofen (e.g., Advil) versus acetaminophen (e.g., Tylenol). A cautious style of driving may promote both wearing seatbelts and staying a safe distance from the car ahead, thereby biasing a comparison of belted and unbelted drivers; however, that cautious style may not bias a comparison of two people in the same car in the same crash, one belted and the other unbelted.

In the examples we have considered, time did not play a central role. In the case of NSAID pain relievers, it was long-term use that mattered. In the case of seatbelts, it was the moment of the crash that mattered. In some situations, time does play a central role; that is, the timing of the event is an essential part of the event.

Expressed differently, over the course of a life, much may be predictable, perhaps rationally planned, perhaps pointlessly habitual, perhaps neurotically determined, or perhaps corralled by social conventions or politics that punish the smallest deviation. Still, in such a life there may be a brief moment when a fateful choice between two very different paths is decided by little more than luck. As the proverb says, "For want of a nail, the horseshoe was lost, then the rider, the battle, and the kingdom." In the gentle language of nonlinear dynamics, there may be "sensitive dependence upon initial conditions," with the trivial deciding between the fortuitous and the catastrophic.

Isolation refers to sifting a large collection of data to collect these rare, brief moments in which chance plays a decisive role in shaping a life. Here, the temporal structure of a life is important and must be respected. We seek two or several people whose lives appeared similar up to a certain moment in their lives, a moment that changed everything for one of these people, that person having been selected by little more than luck. We seek this comparison repeatedly, creating many such small comparisons or matched sets. Inside any one matched set, lives before the fateful event appeared similar; then, for one person, everything changed. In different matched sets, the situations are typically very different, with a different temporal structure and history.

Expressed in slightly technical terms, isolation combines risk-set matching from Chapter 11 and differential effects from this chapter. Treatments occur or not at varied points in time, and risk-set matching pairs people who were

similar up to that time, making no use of information about the future be-
yond that time. At that fateful time, two alternative paths or treatments
become momentarily available—a differential comparison—and a matched
pair or set contains people who took different paths. Different matched sets
have different histories, indeed histories of different lengths, containing in-
formation of qualitatively different types. The cancellation of generic unob-
served biases that we have seen repeatedly in differential comparisons does
occur here, but only at a single instant, and isolation structures a study to
focus on many differential comparisons at their decisive moments.[23]

Mothers, Children, and Careers

What is the effect of having a child on a mother's career? This question, which
can seem to make sense, is constantly at risk of collapsing into a circular
incoherence. For many if not most mothers, family, children, and career are
different parts of one plan, not events that cause one another. One mother
may leave school with a high school degree and marry, then have many
children and no career outside the home, in perfect conformity with her plan.
Another mother may delay marriage and children until age 40, after she has
completed her PhD in biochemistry and her MBA, then raised adequate ven-
ture capital for her start-up, in perfect conformity with her plan. It makes
no sense, in these two cases, to ask about the effects of children on careers
or careers on children.

   And yet, as creatures, we were not intended to have perfect plans for our
fertility. The perfect plan for an individual might produce too few children
for the species, something that evolution does not tolerate for long in an
ongoing species. The random elements that intrude to disrupt our perfect
plans are the building blocks for observational studies.

   What is the effect of having a child on a mother's career? The question
becomes coherent at brief moments when chance intervenes.

   An interesting study by Joshua Angrist and William Evans used U.S.
Census data and two aspects of fertility that are largely determined by chance,
namely, twins and the gender of children.[24] A mother who has twins rather
than a single child may end up with one more child than she had planned.
In the United States there is a tendency for families to prefer having both a
son and a daughter, and if their first two children have the same sex, they

are more likely to try again. These random elements begin to play a role during the second pregnancy, and play a role in subsequent pregnancies. For instance, a mother who wanted two children may end up with three because she had twins during her second pregnancy, or a mother who wanted three children may end up with four because she had twins during her third pregnancy. Similarly, a mother who wanted two children, a son and a daughter, may have a third child because her second pregnancy gave her a second son. Angrist and Evans demonstrated that these patterns are present in Census data; for instance, women are more likely to have a third child if the first two children have the same gender.

Whether or not a woman is pregnant at a certain age, having completed a certain level of education may indicate much about her plans and aspirations. Compare a woman who is pregnant at age 19 with one who is not, and it would not be surprising to find that they have different plans for family, education, and career. That comparison could be very biased by different aspirations that you cannot see in available data. However, given that a woman is pregnant at age 19, whether that pregnancy produces a boy or a girl, a twin or a single birth, is mostly luck. There is a brief moment when certain specific aspects of fertility are just luck, even though most aspects, most of the time, are not just luck. Isolation refers to focusing on the moment and aspect that are determined by luck, to the exclusion of everything else.

A recent reanalysis of Angrist and Evans's 1980 Census data by José Zubizarreta and colleagues focused on the two extreme cases that are determined by luck.[25] In one case, luck is most likely to have given a mother one more child than she originally planned, and in the other case luck is least likely to have done so. A second, third, or fourth pregnancy that produces twins may have given a mother one extra child she did not plan. A second, third, or fourth pregnancy that results in at least one child of each sex is least likely to encourage a mother to have an additional child out of dissatisfaction with the genders of her children. Given that two women are pregnant with their second child at the same age, it is mostly luck whether the first has twins and the second has a child with a different sex than her first child.

Matched sets of six mothers were formed. Each set contained one mother who had twins. Each set contained five mothers who had a single child such that the mother now had at least one boy and one girl. First, matched sets

were created from second pregnancies. Mothers were matched for their characteristics prior to their second pregnancy, the calendar year of the first pregnancy, their age at first pregnancy, their age at second pregnancy, their education level at first and second pregnancies, their race and ethnicity, their region of the United States, and other variables. Mothers who were not used in this first match might be matched at the third pregnancy, now matching for the history leading up to the third pregnancy. Mothers not used in these two matches might be matched at the fourth pregnancy, now matching for the history leading up to the fourth pregnancy. So a total of 5,040 matched sets were formed of six similar women at a similar moment in their lives, a rare moment when luck played a role in deciding the number of children each woman would have.

Although it was not true in every single case, in aggregate the women with twins typically had one more child on Census day 1980 than did their matched controls. This was not true in every case for the simple reason that the women were free to have as many children as they wanted, but typically in this setting the twin added one child to the family's size on Census day, and did so by luck. Importantly, on Census day, the women with twins were working somewhat fewer hours a year, but the difference in hours worked was actually quite small. Angrist and Evans had reached similar conclusions using a different methodology.

Compare women with different numbers of children, and you are mostly comparing women who wanted different lives and are living the lives they deliberately chose. Isolation means sifting a large collection of data to construct a smaller natural experiment in which a treatment that is typically assigned in a biased fashion has, in brief moments and rare situations, been determined mostly by luck.

Taking Stock

A generic unobserved bias is an unobserved covariate, u_i, which promotes several or many related behaviors in a parallel manner. We often speak of a disposition or a penchant to behave in a certain way, and this grouping of related behaviors under a single disposition gains mathematical expression in a generic unobserved bias. Generic biases can sometimes be broken up, eliminated, or reduced by differential comparisons that overcompensate for

observed behaviors in an effort to adequately compensate for an underlying disposition. Differential comparisons concern the effects caused by giving one treatment in lieu of another treatment. Care is needed in the selection of differential comparisons if they are to be informative. Generic biases are one type of bias, and more than one bias may be present. Even if a generic bias is eliminated, a sensitivity analysis is needed to address remaining differential biases that promote just one of several behaviors.

Instruments

What Are Instruments?

Haphazard Nudges amid Biased Treatment Assignment

When people receive treatments in a biased manner, they may at times be pushed toward one treatment or another by a force that is haphazard—effectively random. Expressing the same thought in different words, treatment assignment may have an aspect that is deliberate or systematic, and another aspect that is little more than luck. We describe our lives, our big choices, in narratives that blend conviction and luck, articulate plan and stumbling, seizing opportunity and being in the wrong place at the wrong time. We may be less than candid with ourselves and others in these narratives, but few would deny a role to luck.

An instrument is a random push toward receiving one treatment rather than another, a push that is without consequences for outcomes unless it succeeds in changing the treatment received.[1] So an instrument, to be an instrument, must meet several standards. It must be random: there must be no reason it pushed you and not the person standing next to you. It must push: although some, perhaps many, resist the push, others do enter treatment because they were pushed. Finally, the push must be inconsequential

unless it changes the treatment received. This third requirement is called the "exclusion restriction." Instruments are rare things. Are they useful?

Proximity to a Capable Hospital

A heart attack or acute myocardial infarction (AMI) is typically caused by a narrowing and subsequent blockage of an artery that supplies blood to the heart. Treatment for an AMI may include a physical intervention to open or bypass the blocked artery, performed by a cardiologist or cardiac surgeon. Alternatively, treatment may emphasize drugs that attempt to dissolve the blockage or thin the blood. Which treatment is given to which patient? Many clinical factors enter into a thoughtful decision, but there is one factor that is comparatively haphazard, with no clinical purpose. If nearby hospitals lack the staff and facilities to open or bypass a blocked coronary artery, a patient rushed to the nearest hospital is much more likely to be treated with drugs. In this situation, and many similar situations in health outcomes research, some measure of proximity to a hospital capable of intervening with a particular treatment is used as an instrument for receipt of that treatment.

In a general discussion of the use of instruments in health outcomes research, Joseph Newhouse and Mark McClellan first show that patients whose blocked coronary arteries were opened or bypassed were very different from patients treated with drugs, but patients living near hospitals capable of opening or bypassing blocked arteries were not visibly different from patients living far from such hospitals.[2] For instance, people treated solely with drugs were substantially older, but a patient's age was unrelated to proximity to a capable hospital. In effect, they presented two tables analogous to Table 1.1: one very biased table comparing patients based on how they were treated and the other without visible bias comparing patients based on where they lived. Again, we are less worried about age, which is measured and can be controlled, and we are more worried about some other covariate that was not measured; however, the pattern we see for age might reasonably be expected for many clinically relevant covariates, measured or not. Newhouse and McClellan argued, plausibly if not conclusively, that their measure of proximity to a capable hospital satisfies the conditions that define an instrument. That is, they argued that (i) proximity to a capable hospital tells

you nothing of clinical relevance about the patient; (ii) proximity pushes some people toward a treatment that they would not otherwise receive; and (iii) proximity affects patient outcomes only indirectly by switching the treatments some patients receive. These are assumptions about how the world works, and they might be wrong, but they are not implausible assumptions, and there is no obvious sign in their data indicating that the assumptions are false. It is not easy to check whether a proposed instrument actually is an instrument, as it is not easy to check ignorable treatment assignment, but an elaborate theory can help in the former task as it helped in the latter in Chapter 7.[3]

Distance to a capable hospital is widely used as an instrument in the study of health outcomes.[4] Another approach focuses on habits of particular physicians or hospitals.[5] For instance, one physician typically prescribes an older, off-patent, less-expensive statin, where a different physician typically prescribes a newer, more expensive, patented, brand-name statin. Can this prescribing preference be an aid in comparing the effectiveness of these two statins in terms of cardiac outcomes? Are prescribing habits an instrument? Perhaps, or perhaps not. Harry received the second statin because he found the first physician disagreeable, and so he switched to the second physician. Harry found the first physician disagreeable because she kept telling him to quit smoking and lose weight, where the second physician stuck to prescriptions. Perhaps the statins are equally effective, but for patients more sensible than Harry the first physician is more effective. If so, this would violate one of the assumptions that defines an instrument, the exclusion restriction, namely, that physicians affect cardiac outcomes only indirectly through their choice of statin. It is not easy to check whether a proposed instrument actually is an instrument.

Lotteries with Loopholes

In Chapter 6, lotteries were a source of natural experiments. Some lotteries do not determine treatment assignment; they have loopholes. Some lotteries open or close doors, but do not force people through open doors or close off all entrances. Some lotteries push people toward a treatment without deciding the treatment. Lotteries of this sort may yield instruments rather than natural experiments.

The Vietnam War draft lottery was effectively random, and it pushed some people into military service who would not otherwise have served. However, many people volunteered without being drafted, and some people found ways around the draft. So the draft lottery was a randomized push in the direction of military service. Few people are interested in the effect of the lottery itself, but many people are interested in whether military service is a benefit or harm to income or health. Joshua Angrist and colleagues used the draft lottery as an instrument for military service in studies of the effects of military service on income and health.[6]

Randomized Encouragement Experiments

What Are Encouragement Experiments?

The clearest example of an instrument is the randomized encouragement experiment proposed by Paul Holland.[7] In some randomized encouragement experiments, the conditions required of an instrument are simply true. In these cases, we can set aside the inevitable worries about whether a purported instrument is actually an instrument, and focus instead on what we would do with an instrument if we had one.

A paired randomized encouragement experiment closely resembles a paired randomized experiment in all respects but one. Pairs are matched before treatment based on observed covariates, x_i, and one person in each pair is picked by random numbers for the treatment, the other receiving the control. In an encouragement experiment, the treated individuals are encouraged to do something, perhaps study harder for an academic test, or quit smoking, or have software installed on their cell phones with encouragement to use it. Controls are not encouraged. For instance, the software might offer Harry the opportunity to conveniently and reliably keep track of alcoholic drinks consumed each day, provided that Harry enters this information as drinks are consumed. Some treated people heed the encouragement and do what they are encouraged to do, but many ignore encouragement. The outcome under study might be academic test performance, or a measure of lung function, or arrests for drunk driving.

There are two very different questions. What is the effect of encouragement? What is the effect of doing what you are encouraged to do? It may be

that studying for a test massively improves test performance, but encouragement to study improves test performance only slightly, because most people ignore the encouragement and do not study. It may be that quitting smoking massively improves lung function, but encouragement to quit smoking has only slight effects, because most people ignore the encouragement and do not reduce cigarette consumption. It may be that keeping track of alcoholic drinks as you consume them greatly reduces arrests for drunk driving, but receiving a piece of software has only slight effects because most people decline to use it.

Estimating the effect caused by encouragement is straightforward in a randomized encouragement experiment. The situation is no different from any paired randomized experiment. People were randomly assigned to encouragement or control, end of story.

The problem comes if we are interested in the effects, not of encouragement, but rather of the thing one is encouraged to do. One might wish to say to someone, "I know this is hard for you to hear or act upon, but if you could hear it and act upon it, then you could massively improve your life—look, I have experimental results that show this conclusively." The problem is that the randomized experiment does not demonstrate this in a simple way: it shows a small effect of wise advice, not a massive effect of heeding wise advice.

We cannot simply compare people who heed good advice and people who ignore it. That is not a randomized comparison. People who study for exams when encouraged to do so are different from people who ignore such sensible advice. The same is true for people who quit smoking or monitor their drinking. Perhaps one of the largest and most consequential distinctions among people is that some can heed wise advice, and others cannot. If the goal is to answer the second question—the effects of heeding advice, rather than the effects of having it tossed in your direction—then the encouragement design has randomized something, but not the right thing. Is there some other way to estimate the effects of heeding good advice?

The Instrumental Estimate Described Informally

A formal or mathematical description of the instrumental estimate is coming soon. First, consider a simple case described informally in English.

Consider a paired randomized experiment on smokers in which treated smokers are encouraged to quit, and nothing is done to controls. Suppose that half of the encouraged smokers quit smoking, half do not change their smoking behavior, and no one in the unencouraged control group quits smoking. Here, I am supposing a simple situation so that the instrumental estimate has a simple description. Suppose that average lung function improves by one unit in the encouraged group compared with the unencouraged group. In this case, the logic of Chapter 2 would lead us to estimate the average effect of encouragement as one unit. Encouragement works: it improves lung function by one unit on average. Perhaps some people improve by more than one unit and others improve by less, but the average improvement is one unit. This would be a good estimate of the average effect caused by encouragement, because encouragement was randomly assigned to smokers. In a sense, there is no better estimate of the effect of encouragement. It is the relevant estimate for someone who is trying to decide whether to encourage people to quit smoking. That individual can encourage or not, and the average effect of encouraging is one unit.

However, we want the effect of quitting smoking on lung function, not the effect of encouragement to quit smoking. If I am trying to decide whether I should quit smoking, I want to know the effect of quitting—after all, I am deciding whether to quit, not whether to encourage someone else to quit. Presumably, the effect of quitting is larger, better, than one unit. After all, the one-unit improvement is derived from an average over the encouraged group, and half the encouraged group did not change their smoking behavior. Those slackers pulled the average down. Presumably, the half of the encouraged group that did not change its smoking behavior received no benefit from encouragement it did not heed. The only way unheeded encouragement could improve lung function is if loud arguments with doctors or nurses clear toxins from your lungs. Presumably, the entire benefit is concentrated in the half of the encouraged group that quit smoking. Presumably, the one-unit average gain is half "no gain" for the slackers and half whatever gain comes from actually quitting. Hence, the gain from quitting is estimated to be two units.

Notice that the instrumental estimate starts with the effect of encouragement, but it attributes the entire effect of encouragement to the subset of people who heed encouragement. The justification for this attribution is an assumption that is possibly false, namely, the exclusion restriction. The

exclusion restriction says that encouragement affects lung function only by changing smoking behavior. The popular saying "no pain, no gain" is the exclusion restriction stated in words of one syllable. The exclusion restriction is very plausible in this example; it is not actually plausible that loud arguments clear toxins from your lungs.

Is a two-unit gain a good estimate of the effects of quitting? We are confident that the encouragement was randomized. We see in the data that encouragement caused half of those encouraged to quit. The first two aspects of an instrument are in place: the push was randomized, and the push really was a push. The third condition is the exclusion restriction: encouragement must affect lung function only by changing smoking behavior. If the exclusion restriction were true—it certainly sounds plausible enough here—then a two-unit gain would indeed be a good estimate of the effect of quitting smoking.

Effect Ratios

An effect ratio in a randomized experiment is simply the average treatment effect for one outcome divided by the average treatment effect for another outcome. In principle, effect ratios are not new, and they need not have anything to do with instruments; rather, we take two familiar things and divide one by the other. Let us say that more carefully.

Suppose that a randomized experiment has two outcomes. The first outcome is simply (r_{Ti}, r_{Ci}), as before, with observed value $R_i = r_{Ti}$ if individual i received treatment with $Z_i = 1$ or $R_i = r_{Ci}$ if individual i received control with $Z_i = 0$. As in earlier chapters, for outcome (r_{Ti}, r_{Ci}), let us write $\delta_i = r_{Ti} - r_{Ci}$ for the effect difference for person i and $\bar{\delta} = (1/I)(\delta_1 + \cdots + \delta_I)$ for the average treatment effect. So far, there is nothing new here, just repetition of notation from Chapter 2. For a second outcome, we need a second symbol, but aside from the second symbol everything is exactly parallel with the first outcome. Let us write (s_{Ti}, s_{Ci}) for the second outcome, with observed value $S_i = s_{Ti}$ if individual i received treatment with $Z_i = 1$ or $S_i = s_{Ci}$ if individual i received control with $Z_i = 0$. For the second outcome, let us write $\eta_i = s_{Ti} - s_{Ci}$ for the effect on person i and $\bar{\eta} = (1/I)(\eta_1 + \cdots + \eta_I)$ for the average treatment effect.

An effect ratio is the ratio of two average treatment effects, say $\bar{\delta}/\bar{\eta}$. Implicitly, we are assuming $\bar{\eta} \neq 0$ to avoid a division by zero, but let us worry

about that later. Reasoning as in earlier chapters, it is easy to estimate an effect ratio in a randomized experiment: (i) estimate $\bar{\delta}$ by the mean of R_i in the treated group minus the mean of R_i in the control group; (ii) estimate $\bar{\eta}$ by the mean of S_i in the treated group minus the mean of S_i in the control group; and (iii) divide the first estimate by the second. So there is nothing special about estimating effect ratios in randomized experiments, provided that the denominator is not close to zero.

Chapter 3 tested hypotheses and built confidence intervals in randomized experiments. How would we test hypotheses about the value of an effect ratio, $\bar{\delta}/\bar{\eta}$? Consider the hypothesis H_ρ that $\bar{\delta}/\bar{\eta}$ is some specific number ρ, perhaps $\rho = 17$, or any other specific number. As with estimating $\bar{\delta}/\bar{\eta}$, it will turn out with a little algebra that there is nothing special about testing the hypothesis H_ρ about the value of an effect ratio—it is the same as testing that a certain average treatment effect is zero. The algebra is easy, so I will present it, but the algebra is not needed later, so you can skip the rest of this paragraph if you are so inclined. If H_ρ were true, then $\rho = \bar{\delta}/\bar{\eta}$, so that

$$\rho = \frac{\delta_1 + \cdots + \delta_I}{\eta_1 + \cdots + \eta_I}$$

or equivalently $\rho\eta_1 + \cdots + \rho\eta_I = \delta_1 + \cdots + \delta_I$, or equivalently $0 = \delta_1 - \rho\eta_1 + \cdots + \delta_I - \rho\eta_I$. So H_ρ is equivalent to the hypothesis that the average effect of encouragement on the quantity $\delta_i - \rho\eta_i$ is zero. Of course, $\delta_i - \rho\eta_i$ is by definition $(r_{Ti} - r_{Ci}) - \rho(s_{Ti} - s_{Ci})$ or the difference in effects for the outcome $(r_{Ti} - \rho s_{Ti}, \ r_{Ci} - \rho s_{Ci})$, whose observed value is $R_i - \rho S_i$. That is, the hypothesis H_ρ that $\rho = \bar{\delta}/\bar{\eta}$ is the same as the hypothesis of zero average effect for the observed outcome $R_i - \rho S_i$.

Keep in mind that the null hypothesis H_ρ stipulates a numeric value for ρ, say, $\rho = 17$, so in testing H_ρ using $R_i - \rho S_i$ we use the value of ρ stipulated by the null hypothesis, say, $R_i - (17 \times S_i)$. We test H_ρ by testing that $R_i - \rho S_i$ has the same mean in the treated and control groups, which is hardly rocket science. Testing hypotheses about an effect ratio, $\bar{\delta}/\bar{\eta}$, is transformed by two lines of algebra into a standard problem of testing hypotheses about one average treatment effect.[8]

Because, by definition, a 95% confidence interval is the set of hypotheses not rejected at the 0.05 level by a test, a confidence interval for $\bar{\delta}/\bar{\eta}$ is obtained by testing each possible value of ρ and retaining the values not rejected.[9]

So there is nothing special about hypothesis tests and confidence intervals for effect ratios in randomized experiments—it is the old machinery plus two lines of algebra. In particular, the exclusion restriction plays no role in inference about effect ratios.

So far, effect ratios are one topic, and instruments are another. Are they related?

Effect Ratios and the Exclusion Restriction

Consider, again, the case of encouragement to quit smoking. The first outcome, (r_{Ti}, r_{Ci}), is the primary outcome—say, the effect of encouragement to quit smoking on lung function. The second outcome, (s_{Ti}, s_{Ci}), indicates whether encouragement to quit smoking caused individual i quit smoking: 1 for quitting, and 0 for no change in smoking behavior. A person i with $(s_{Ti}, s_{Ci}) = (0,0)$ and $\eta_i = s_{Ti} - s_{Ci} = 0 - 0 = 0$ would not quit whether encouraged or not. A person i with $(s_{Ti}, s_{Ci}) = (1,1)$ and $\eta_i = s_{Ti} - s_{Ci} = 1 - 1 = 0$ would quit spontaneously whether encouraged or not. A person i with $(s_{Ti}, s_{Ci}) = (1,0)$ and $\eta_i = s_{Ti} - s_{Ci} = 1 - 0 = 1$ would quit if encouraged, but would not quit without encouragement; that is, a person who would be caused to quit smoking by being encouraged. A slightly perverse person i with $(s_{Ti}, s_{Ci}) = (0,1)$ and $\eta_i = s_{Ti} - s_{Ci} = 0 - 1 = -1$ would not quit if encouraged to quit, but would quit spontaneously if not encouraged. Also, a person might cut back on smoking without quitting. Tidy conclusions are obtained by assuming there are no such perverse people and no one cuts back—everyone quits or does nothing; so to see these tidy conclusions we will make these unneeded assumptions in this section only.[10]

Angrist, Imbens, and Rubin used the names "never-taker," "always-taker," "complier," and "defier" for people with $(s_{Ti}, s_{Ci}) = (0,0)$, $(s_{Ti}, s_{Ci}) = (1,1)$, $(s_{Ti}, s_{Ci}) = (1,0)$, and $(s_{Ti}, s_{Ci}) = (0,1)$, respectively.[11] These terms indicate the circumstances under which person i changes behavior in the direction that is encouraged: never, always, only if encouraged, or only if not encouraged. Assuming that there are no perverse people who always do the opposite of what they are encouraged to do is the same as assuming there are no defiers.

In this context, the exclusion restriction again says no pain, no gain: encouragement affects your lung function only if encouragement causes you to quit. That is, encouragement only affects compliers. By definition, encour-

agement does not cause quitting if $0 = \eta_i = s_{Ti} - s_{Ci}$. By definition, encouragement does not cause an improvement in lung function if $0 = \delta_i = r_{Ti} - r_{Ci}$. The exclusion restriction says encouragement to quit fails to cause an improvement in your lung function unless it causes you to quit; that is, the exclusion restriction says $0 = \delta_i = r_{Ti} - r_{Ci}$ whenever $0 = \eta_i = s_{Ti} - s_{Ci}$. Encouragement can have an effect on lung function with $\delta_i \neq 0$ only if encouragement had an effect on smoking, $\eta_i \neq 0$. Now comes an important step. The exclusion restriction means that the average effect of encouragement, $\bar{\delta}$, is an average of many zeros for people who would not heed encouragement and some possibly nonzero δ_i's for compliers who do heed encouragement.

What does the exclusion restriction mean for an effect ratio, $\bar{\delta}/\bar{\eta}$? Because $\bar{\delta}$ and $\bar{\eta}$ are each averages of I terms, $I\bar{\eta}$ and $I\bar{\delta}$ are each sums of I terms. For instance, $I\bar{\eta}$ is the number of people who would quit smoking only if encouraged—it is the number of compliers, those with $(s_{Ti}, s_{Ci}) = (1,0)$ and $\eta_i = 1$, because we have assumed there are no defiers with $\eta_i = -1$. The exclusion restriction means that $\delta_i = 0$ whenever $\eta_i = 0$, so it follows that $I\bar{\delta}$ is the total of some $\delta_i = 0$ for people who do not heed encouragement and some possibly nonzero δ_i's for compliers who do heed encouragement. It follows that the effect ratio, $\bar{\delta}/\bar{\eta} = (I\bar{\delta})/(I\bar{\eta})$, is the total of δ_i's for people who heed encouragement divided by the number of people who heed encouragement. The slackers who do not change their behavior in response to encouragement—the slackers who brought the average down—are not part of $\bar{\delta}/\bar{\eta}$. In other words, $\bar{\delta}/\bar{\eta}$ is the average effect of encouragement on compliers.[12] Stated briefly if somewhat imprecisely, $\bar{\delta}/\bar{\eta}$ is the effect of changing behavior when encouraged to do so, the effect of quitting on lung function, just what we wanted.

Let us compare this general finding with the informal numerical example that preceded it. In the numerical example, encouragement caused half of those encouraged to quit, so $\bar{\eta} = 1/2$. The average effect of encouragement on lung function was $\bar{\delta} = 1$. So the effect ratio was $\bar{\delta}/\bar{\eta} = 1/(1/2) = 2$. The effect ratio, $\bar{\delta}/\bar{\eta}$, equals the average effect of quitting on people who quit only when encouraged to do so, provided that the exclusion restriction holds—that is, provided that encouragement only affects people who heed it.

The effect ratio, $\bar{\delta}/\bar{\eta}$, makes perfect sense whether or not the exclusion restriction holds. Without the exclusion restriction, the effect ratio, $\bar{\delta}/\bar{\eta}$, is an accounting parameter, how much of this for how much of that. For instance, in the numerical example in the previous paragraph, $\bar{\delta}/\bar{\eta} = 2$ means an

average of two units improvement in lung function for each person caused to quit smoking by encouragement. The accounting interpretation is fine so far as it goes, but it makes no claim that the causal effect in the numerator has anything to do with the causal effect in the denominator. With the exclusion restriction, the effect ratio, $\bar{\delta}/\bar{\eta}$, is a causal parameter, the average benefit that accrues to quitters who quit only when encouraged to do so. That is, with the exclusion restriction, the effect ratio, $\bar{\delta}/\bar{\eta}$, is the average effect of the change in behavior for compliers.

Weak Instruments

Three conditions are required if encouragement is to be an instrument: encouragement must be randomized, it must push some people to change their behavior, and encouragement affects outcomes only to the extent that encouragement alters behavior. The first condition is guaranteed in a randomized encouragement experiment, and the third condition is the exclusion restriction. What about the second condition? What if the push is not much of a push? What if lots of encouragement to quit smoking yields only a few quitters? What if it yields none at all?

An instrument is said to be weak if its push is not much of a push. If almost everyone ignores encouragement, the instrument is weak.

If encouragement had no effect on anyone's smoking behavior, then the average effect of encouragement on smoking behavior would be zero, $\bar{\eta} = 0$, and the effect ratio, $\bar{\delta}/\bar{\eta}$, would entail a division by zero. In this case, there is something defective about the parameter itself: $\bar{\delta}/\bar{\eta}$ is the average effect of quitting on people who quit only when encouraged to quit, in a world in which no one actually does quit only when encouraged to quit. If $\bar{\eta} = 0$, then something is wrong with $\bar{\delta}/\bar{\eta}$.

If $\bar{\eta} \neq 0$ but $\bar{\eta}$ is near zero, then the parameter $\bar{\delta}/\bar{\eta}$ is well defined, but estimates of $\bar{\delta}/\bar{\eta}$ can be very unstable. That is, if very few people quit smoking when encouraged to do so, it becomes very difficult to estimate the effect of quitting on people who quit when encouraged. With appropriate methods in a randomized encouragement experiment, a weak instrument will simply yield a long confidence interval for $\bar{\delta}/\bar{\eta}$.[13] Alas, several of the most popular methods, including two-stage least squares, are not appropriate in this sense,

and even their confidence intervals give the wrong answer with a weak instrument.[14]

Instruments in Randomized Clinical Trials: Noncompliance with Assigned Treatment

A patient is accepted into a randomized clinical trial only if the patient signs a document indicating that the patient has been informed about how the trial will proceed and the trial's risks, and voluntarily agrees to participate in light of this information. It would be a substantial breech of ethics and law to trick or coerce someone into participating in a clinical trial. Nonetheless, having provided informed consent, a person is a person, and people change their minds. Halfway through treatment, Harry decided to move from Dallas to Cincinnati to be close to his daughter, and the investigators in Dallas found it impractical to continue to treat Harry in Cincinnati, though they continued to monitor Harry's health outcomes. Every time Sally took the assigned medication, she became extremely nauseous, and after the third dose she had had enough, but the investigators continued to monitor Sally's health outcomes.[15] Situations like this are called "noncompliance" with the assigned treatment.

A randomized experiment with noncompliance is an encouragement experiment. Randomization encouraged one person to take treatment, another to take control, but not everyone did what they were encouraged to do. When there is noncompliance, one analysis is invariably done, called the intent-to-treat analysis, comparing everyone randomly assigned to treatment and everyone randomly assigned to control, ignoring whether they complied or not. Harry's outcomes are credited to the group to which he was assigned, and Sally's outcomes are credited to the group to which she was assigned, even though Harry wandered off to Cincinnati and Sally took only three of her 30 doses of medication. This intent-to-treat analysis is identical to studying the effects of encouragement in any encouragement experiment. In Sally's case, it is important that not everyone can tolerate the medication, that noncompliance is a feature of the drug, not of the randomized trial. In Harry's case, if the drug turns out to be a big success, someday it will be available in Cincinnati too, so Harry's noncompliance is a feature of the

randomized trial, not of the drug. In any event, the intent-to-treat analysis refers to the effect of encouragement.

The intent-to-treat analysis can be, and often is, supplemented with an instrumental analysis that tries to estimate the effect of the drug on people who take it. With appropriate methods, the two analyses provide compatible answers: the drug can only work if encouragement to take it works, but the drug has a larger effect if you take it than if you do not. That is, the null hypothesis of no effect of encouragement must be rejected before saying that the drug has any effect at all, but the instrumental estimate will attribute the effects of the drug to the subset of people who take it, judging the drug to be more effective for compliers than for a mixed group of compliers and noncompliers.[16]

How Do Observational Studies with Instruments Differ from Encouragement Experiments?

The key element in a randomized encouragement experiment is that the treatment, encouragement, is randomized using random numbers. Estimating the effect caused by encouragement is straightforward because encouragement was randomized. The one complication occurs when we try to estimate the effect, not of encouragement, but of actually doing what you were encouraged to do.

In an observational study with a parallel structure, encouragement itself is not randomly assigned. Estimating the effects of encouragement involves dealing with nonrandom treatment assignment, the problems discussed in earlier chapters. On top of these problems, the investigator wants to estimate the effect, not of encouragement, but of what you were encouraged to do.

Commonly, when instruments are used in observational studies, there is a hope, perhaps a realistic hope, that the assignment of encouragement is more nearly random than the assignment of the thing that is being encouraged. This was the point made by Newhouse and McClellan: how a heart attack is treated does not look at all random, but how close you live to a hospital that can perform coronary bypass surgery might be nearly random. It was also the point made by Angrist and colleagues: the Vietnam War draft lottery delivered randomized encouragement to serve in the military, but volunteers and draft-dodgers meant that military service was not random.

Perinatal Care for Premature Infants

Do High-Level Neonatal Intensive Care Units Save the Lives of Premature Infants?

Hospitals vary in their abilities to care for premature infants. The American Academy of Pediatrics grades neonatal intensive care units (NICUs), with high grades assigned to hospitals with greater capabilities. Here, a high-level NICU will refer to grades 3A and above, and the low level will refer to grades 1 and 2. Do high-level NICUs save more lives?

There is a problem with comparing infant survival at high- and low-level NICUs. A mother whose pregnancy is at risk is likely to be told that she should deliver at a more capable hospital, whereas a mother whose pregnancy exhibits little or no apparent risk may be told that the nearest hospital is fine. Indeed, a perfectly rational process called "regionalization" tries to make this happen, to direct mothers to appropriate hospitals. If this rational process worked perfectly, the most capable hospitals would see the most challenging patients. Rationality can be a good thing in health care. Too much rationality can be a problem for an observational study that is seeking random treatment assignment; for this task, erratic, haphazard, and eccentric are preferred to rational. Is it fair to compare the performance of high-level NICUs with challenging pregnancies to the performance of low-level NICUs with routine pregnancies?

A Matched Comparison in Which Distance Discourages Use of High-Level NICUs

A study by Michael Baiocchi and colleagues compared mothers of premature infants in Pennsylvania who lived near a hospital with a high-level NICU to mothers who lived near a hospital with a low-level NICU.[17] The study formed 49,587 pairs of mothers with similar health insurance, socioeconomic status, age, education, race, and parity, with prenatal care beginning in the same month, whose babies were of similar gestational age and birth weight.[18]

By construction, one mother in each pair lived close to a high-level NICU and the other lived far away, in the following sense. The excess travel time is the travel time to the nearest hospital with a high-level NICU minus the

travel time to the nearest hospital. It answers this question: How much longer would it take to reach a hospital with a high-level NICU, as opposed to delivering at the nearest hospital? For mothers near a high-level NICU, the average excess travel time was under one minute; that is, either the closest hospital had a high-level NICU, or there was one just a little farther away. For mothers far from a high-level NICU, the average excess travel time was 35 minutes. Keep in mind that 35 minutes is the excess travel time, not the total travel time; that is, total travel time to a hospital with a high-level NICU is travel time to the nearest hospital plus the excess travel time.

The hope is that, having matched for many covariates, there is nothing special about mothers who live near or far from a hospital with a high-level NICU. The hope is that living far from a hospital with a high-level NICU discourages delivery at such a hospital but is otherwise without consequence for infant survival. The hope, perhaps realistic, is that the 49,587 pairs resemble a paired randomized encouragement experiment, in which distance encouraged some mothers of premature infants to deliver at high-level NICUs and others to deliver at low-level NICUs. Remember that an instrument is a device intended to separate a random aspect of treatment assignment from a biased aspect, and to make effective use of the random aspect, leaving the biased aspect untouched. Repeating for emphasis what has been said several times: it is encouragement—near or far from a high-level NICU—that, one hopes, is nearly random; however, where mothers actually choose to deliver may be biased, not at all random. This hope that encouragement is random within each pair might be realized, yet when mothers do not do what distance encourages them to do, then they may have good reasons for doing what they did instead. Randomized encouragement is useful and compatible with behavior that ignores encouragement in a biased manner.

Do High-Level NICUs Save Lives?

There were 49,587 pairs of similar mothers of premature infants, one who lived close to a high-level NICU, another who lived much farther away. Within these pairs, the mother near a high-level NICU had an excess travel time of, on average, under a minute, often because the nearest hospital had a high-level NICU and the excess travel time to a high-level NICU was zero. On average, the mother far from a high-level NICU would have had to travel

for an extra 35 minutes to reach a hospital with a high-level NICU. In aggregate, 75% of mothers far from a high-level NICU delivered at hospitals with a low-level NICU, but 31% of mothers near a high-level NICU delivered at hospitals with a low-level NICU. So the average effect, $\bar{\eta}$, of encouragement to deliver at a low-level NICU on delivery at a low-level NICU is estimated to be 0.75 − 0.31 = 0.44, or 44% of mothers changing the NICU level of their hospital in response to additional distance. The infant mortality rate was 0.0194 for mothers who lived far from a high-level NICU and 0.0155 for those lived near a high-level NICU, so the estimate of the average effect, $\bar{\delta}$, of encouragement on mortality is 0.0194 − 0.0155 = 0.0039.[19]

These would be reasonable estimates of the effects produced by encouragement or distance if distance were assigned to mothers at random within matched pairs.[20] Of course, $\bar{\delta}$ is not the effect of delivery at a low-level NICU, because some mothers far from a high-level NICU ignored the encouragement provided by distance, as did some mothers who were close to a high-level NICU. The effect ratio, $\bar{\delta}/\bar{\eta}$, is estimated by the ratio of the estimated effects, 0.0039/0.44 = 0.009, or just shy of 1%. This estimate of the effect ratio, $\bar{\delta}/\bar{\eta}$, has both an accounting and a causal interpretation. Assuming just that distance was randomized within pairs, the accounting interpretation is: an average of 0.009 infant lives saved for every mother caused to deliver at a hospital with a high-level NICU by virtue of living near one. If one assumes both the exclusion restriction and that distance was randomized within pairs, then the causal interpretation is: 0.009 infant lives saved per infant among infants born to mothers who shift to a high-level NICU by virtue of living near one. The distinction in English is subtle because English uses causal terminology imprecisely.

The accounting interpretation is comparing the size of two causal effects, but making no connection between them. The causal interpretation is saying that the effect on mortality in the numerator is found only among mothers who would switch in the denominator from a hospital with a low-level NICU to a hospital with a high-level NICU if they lived closer to the high-level NICU. The exclusion restriction denies that there is any benefit for infant survival of living near a hospital with a high-level NICU if delivery does not occur at such a hospital. The exclusion restriction would be false if a baby who was delivered at a hospital with a low-level NICU benefits from having nearby a hospital with a high-level NICU; for instance, this might be the case if it is safe to transfer a baby a short distance, but not a long distance,

to a hospital with a higher-level NICU. The accounting interpretation is available whether or not the exclusion restriction is true.

The estimate of 0.009 for $\bar{\delta}/\bar{\eta}$ comes with a 95% confidence interval of 0.0057 to 0.0123. The confidence interval, like the estimate, depends upon the assumption that encouragement or distance is randomized within pairs. What if that assumption is wrong? The effect ratio, $\bar{\delta}/\bar{\eta}$, would equal zero if living near a high-level NICU had no effect on infant mortality. The one-sided test of this hypothesis of no effect becomes sensitive to unmeasured bias at $\Gamma > 1.23$.[21] The unmeasured bias, u_i, mentioned here is one that connects where a mother lives to infant mortality, in pairs that were matched for many observed covariates x_i. Recall from Table 9.1 that a bias of $\Gamma = 1.25$ would mean an unobserved u_i that doubled the chance of living far from a high-level NICU and doubled the chance of infant death.[22]

In brief, there is some evidence that delivery of a premature infant at a hospital with a high-level NICU reduces mortality when compared with delivery at a hospital with a low-level NICU. The estimate of $\bar{\delta}/\bar{\eta}$ is just shy of 1%, which is a substantial fraction of the relatively low average mortality of these infants. The results are insensitive to small unobserved biases connecting mortality with where mothers lived after adjusting for several measures of socioeconomic status and the degree of prematurity.

Strengthening Instruments

Strengthening Instruments in the Study of Neonatal Intensive Care Units

The study by Baiocchi and colleagues illustrated the possibility of building stronger instruments. The study as described in the previous section had a stronger instrument with 49,587 matched pairs. A parallel study was conducted with 99,174 matched pairs, more than twice as many. The difference between these two studies, besides the difference in sample size, was the strength of the instrument. To increase the number of pairs to 99,174, weak pairs had to be tolerated. Is it better to have fewer but stronger pairs?[23]

In Pennsylvania, most people live in or near the state's population centers, Philadelphia, Pittsburgh, or Allentown-Bethlehem, with the consequence that most people in the state live near hospitals with high-level

NICUs. People who live far from every hospital with a high-level NICU are less common, and are much less common in certain demographic or socioeconomic groups—that is, certain groups defined by x_i.

The smaller study with 49,587 matched pairs required the difference in excess travel time to be larger, meaning that travel time exerted greater force in determining when a mother delivered. Recall that in this comparison the excess travel time was, on average, under a minute for those near a high-level NICU and 35 minutes for those far away. Increasing the number of pairs to 99,174, meant that the average excess travel time rose to 4.5 minutes for those near a high-level NICU, and dropped to 18 minutes for those far from a high-level NICU. Keep in mind that a mother who must travel an extra 35 minutes beyond the nearest hospital to reach a high-level NICU may already have been required to travel quite a distance to reach the nearest hospital with its low-level NICU; that is, perhaps the total trip to the high-level NICU could be an hour while in labor.

Unsurprisingly, a travel time of 35 extra minutes in the smaller study did more to deter mothers than did a travel time of 18 extra minutes in the larger study. In the larger study, more mothers ignored distance when deciding where to deliver because the smaller distances were less of a deterrent. In other words, the estimate of $\bar{\eta}$ was smaller in the larger study, 0.18 rather than 0.44, and the instrument was weaker. In the smaller, stronger study, 44% of mothers were estimated to switch where they delivered in response to 35 minutes of extra travel, while in the larger, weaker study, only 18% of mothers were estimated to switch in response to 18 extra minutes of travel time. The estimate of the effect of distance on mortality, $\bar{\delta}$, was also smaller, 0.0017 rather than 0.0039, less than half the size. In the larger study, a push that is less of a push yielded a smaller change in behavior with a correspondingly smaller change in mortality. As the sample size got larger, all the interesting parts got smaller.

The estimate of the effect ratio, $\bar{\delta}/\bar{\eta}$, was similar in the larger and smaller study, about 0.009, but the confidence interval was much longer, about 70% longer, in the larger study. That is not a typo. The confidence interval was longer with 99,174 matched pairs than with 49,587 matched pairs. The harm done by weakening the instrument more than offsets the increase in sample size. Additionally, the larger study was sensitive to much smaller biases. The larger study with 99,174 pairs was sensitive to biases of $\Gamma > 1.05$ whereas the smaller study with 49,587 was sensitive to biases of $\Gamma > 1.23$.[24] So, in every

respect, the smaller study produced firmer conclusions, essentially because the instrument was stronger, and the push was more of a push.

An Observational Study Closer to an Experiment and Farther from Pennsylvania

Pennsylvania, as it naturally occurs, is not much of an experiment. Too many people in Pennsylvania live in population centers near high-level NICUs, and too many others live just a bit farther away, with many practical choices of hospitals. The smaller study with 49,587 reshaped Pennsylvania into a better experiment. Half the people in the smaller study had quite a trip to a high-level NICU, and that long trip deterred many people, perhaps 44%, so distance exerted a strong push on where a mother delivered. The smaller study matched for key measures of the degree of prematurity and for demographic and socioeconomic factors, so it seems to all visible appearances to be an equitable comparison. Appearances may deceive, but with findings insensitive to biases of $\Gamma \leq 1.23$, the deception must not be trivially small.

Nonetheless, the smaller study looks less like Pennsylvania and more like a laboratory experiment. Is it a problem that the smaller study looks less like Pennsylvania? It would be a problem if the study were primarily about Pennsylvania, if it were primarily an aid to planning for some office of the state's government. However, the study was never about Pennsylvania. The study was about the effects on infant mortality of delivering at a hospital with a high-level NICU. For that question, the boundaries of the state of Pennsylvania define no population of special interest.

Observational studies of causal effects should resemble experiments, not surveys of administrative jurisdictions.

Weak Instruments Are Always Sensitive to Small Unmeasured Biases

In the two versions of the NICU study, the larger study with a weaker instrument was more sensitive to unmeasured biases than the smaller study with a stronger instrument, even though the estimated effect ratio was 0.009

in both studies. Is this a peculiarity of one study or something to be expected in general?

Theoretical arguments show that weak instruments are always sensitive to small biases from unobserved covariates. A strong instrument may or may not be sensitive to small unmeasured biases. The biases under consideration refer to who is selected for encouragement, not who heeds encouragement. Substantial biases in who heeds encouragement do not, by themselves, create a problem for instruments.

Stated more precisely, the strength of an instrument strongly affects the design sensitivity in Chapter 10. A weak instrument yields a design sensitivity close to 1. A bias that is not much of a bias can explain away the effects of a push that is not much of a push. This remains true as the number of pairs increases, $P \rightarrow \infty$. A strong instrument may have a large or a small design sensitivity; that depends upon the consequences of heeding encouragement.[25]

✳ Aspects of Building Stronger Instruments

In building stronger instruments, one must characterize matched pairs as weak or strong without using the two outcomes, R_i and S_i. One can look at a map and determine that people who live in a certain location must drive a long distance to reach a high-level NICU, but one cannot use the hospital S_i chosen by mother i. One can look at a clock and determine that a mother admitted at a certain hour of night is likely to be discharged after an additional day in the hospital, but one cannot look at the number of days S_i stayed by mother i.[26] One can characterize an institution by how that institution behaves toward people not eligible for the study, but not by how that institution behaves, S_i, toward person i in the study.[27]

Specific matching strategies and algorithms are used to optimize the construction of stronger instruments.[28] By long tradition, these algorithms have the awkward, ugly name "nonbipartite matching," which simply means "not two parts" or "not treated-versus-control." Nonbipartite matching divides one population of people into pairs so that no person appears in more than one pair. The matching algorithm may discard some people, reducing the sample size. When used with instruments, the pairs are as close as possible in terms of observed covariates x, but are constrained to differ substantially

in terms of, say, excess travel time. Changing the constraint on travel time changes the strength of the instrument and the sample size.

Taking Stock

An instrument is a random push to receive a treatment, a push that may not be heeded, a push that can affect outcomes only to the extent that it affects the treatment received. Instruments may be useful when treatment assignment is partly determined by purposes, intentions, or other systematic biases, but also partly determined by luck. An analysis using an instrument tries to exploit the aspect of treatment assignment that is determined by luck. A purported instrument may fail to be an instrument if it is not random, if it is not a push, or if the push has effects besides pushing someone toward treatment. Instruments that meet these requirements are hard to find, and it is difficult to marshal evidence demonstrating that one has been found. True instruments are rare, but the scientific literature is full of purported instruments. An instrument is weak if the push is very gentle, barely altering behavior. Weak instruments are invariably sensitive to small failures of randomness, leading to a preference for strong instruments. A bias that is not much of a bias can explain away the effects of a push that is not much of a push. Certain matching algorithms attempt to construct strong instruments.

Conclusion

John Stuart Mill wanted causal inference to compare identical people under alternative treatments, but this is impossible once one takes account of more than a few of the ways people can differ.[1]

Randomized experimentation is an approach to causal inference introduced by Sir Ronald Fisher. Fisher did not make people the same. Rather, he made the treatment a person received unrelated to everything that makes people different—their genomes, their unconscious thoughts, everything. Fisher's program is easy to implement: you flip a fair coin.

Randomized experimentation is not always practical, and not always ethical. Many treatments cannot be imposed upon people simply to discover the effects they cause. The effects caused by certain treatments must be investigated in observational studies, that is, in nonrandomized studies of treatment effects.

Treatment assignment is ignorable given observed covariates x if (i) every person has a nonzero chance of receiving treatment and of receiving control, and (ii) two people, say, person i and person j, with the same observed covariates, $x_i = x_j$, have the same probability of receiving treatment, $\pi_i = \pi_j$. If treatment assignment were ignorable given x, then appropriate adjustments for x would permit inference about the effects caused by treatments. For instance, if the treatment assignment were ignorable given observed covariates

x, then it would suffice to match people for a single variable, the propensity score, λ_x. The central problem in observational studies is that we have no way to ensure that treatment assignment is ignorable given observed covariates *x.* Most of the controversy that surrounds observational studies reflects concern that treatment assignment is not ignorable given observed covariates *x,* and that correct causal inferences require adjustments for both *x* and a covariate *u* that was not observed.

Natural experiments are attempts to find circumstances in the world that resemble well-designed randomized experiments. In a natural experiment, treatments are not assigned by coin flips, but perhaps they are assigned by something haphazard, aimless, irrelevant. Departures from ignorable treatment assignment resemble descending a gradual slope, not falling off a cliff, and a natural experiment is an attempt to remain close to the summit. Natural experiments often have additional features typical of well-designed experiments, such as sharply distinct treatments that begin abruptly at a well-documented moment.

In the absence of random assignment of treatments, an association between treatment received, *Z,* and outcome exhibited, *R,* is ambiguous, possibly an effect caused by the treatment, possibly a bias in the way treatments were assigned to individuals. Elaborate theories and quasi-experimental devices are two approaches to reducing this ambiguity. Elaborate theories sharpen the description of a causal theory so that it makes more extensive contact with data that we can observe, thereby helping to distinguish treatment effects and biases. Quasi-experimental devices add elements to the design of an observational study, such as multiple control groups and counterpart comparisons. These design elements are intended to eliminate particular rival explanations of an ambiguous association between treatment and outcome.

When applied to practical concerns, every mathematical calculation, including every statistical analysis, depends upon assumptions. What if the assumptions are wrong? A sensitivity analysis introduces quantified violations of an assumption and determines the smallest violation that would alter the conclusion. In an observational study, a sensitivity analysis permits gradually increasing departures from ignorable treatment assignment, thereby determining the magnitude of bias from an unmeasured covariate *u* that would need to be present to alter the study's conclusions. Observational studies vary enormously in their sensitivity to unmeasured bias. Small bi-

ases could radically alter the conclusion of some studies, whereas only large biases could alter the conclusions of other studies. The degree of sensitivity to unmeasured bias is one element in appraising evidence of cause and effect. The degree of sensitivity to unmeasured bias is not indicated by confidence intervals, P-values, or estimates computed assuming ignorable treatment assignment. Rather, a separate analysis is required that permits gradually increasing violations of ignorable treatment assignment.

Having discovered that some observational studies are sensitive to small unobserved biases while other studies are insensitive to fairly large biases, it is natural to ask how we can design observational studies so that they are insensitive to larger biases. The design sensitivity is an aid to answering that question. The design sensitivity is computed from a sampling model that generates data and from particular methods for analyzing these data. The design sensitivity is the limiting sensitivity to unmeasured bias as the sample size increases. Using the design sensitivity, we discover that certain sampling models and methods of analysis yield conclusions that are insensitive to large biases. We then try to find worldly circumstances that produce data similar to the data produced by these sampling models.

Generic biases are unmeasured dispositions u that promote several treatments in a similar way. Unlike most unobserved biases, unmeasured generic biases can sometimes be reduced or eliminated by overcompensation for measured behaviors.

Some treatment assignments have elements that are biased and systematic, other elements that are random. An instrument is a haphazard push or encouragement to receive one treatment rather than another, where encouragement can affect outcomes only if it succeeds in altering the treatment. True instruments are rare. It is difficult to check whether a purported instrument is indeed an instrument. Analysis using an instrument compares encouraged and unencouraged groups, attributing any difference in outcomes between these groups to the people whose behavior was altered by encouragement.

APPENDIX
Bibliographic Remarks

Sir Ronald Fisher invented randomized experimentation in the 1920s and 1930s, and his 1935 book *Design of Experiments* continues to be worth reading.[1] The reasoning in chapter 2 of Fisher's book defines the logic of randomized experimentation and inference in trials such as the ProCESS Trial. David Cox and Nancy Reid provide a modern textbook discussion of the statistical theory of the design of experiments, carefully discussing the logic of randomization and the inferences derived from it.[2] Several articles and books survey the history or practice of randomized experimentation in clinical medicine, education, public policy, criminology and economics.[3]

In the context of randomized experimentation, Jerzy Neyman appears to have been the first person to express causal effects as comparisons of potential outcomes under alternative treatments.[4] In discussing randomized experiments, Fisher's work and Neyman's work were frequently used as a unified whole.[5] In the 1970s, Donald Rubin developed these ideas to clarify the relationship between experiments and observational studies, thereby substantially deepening the ideas and broadening interest in them.[6] For an appreciation of Frederick Mosteller, see the obituary by Ivan Oransky.[7]

Some ideas in Chapter 5 are very old. Direct adjustment or standardization of rates is a part of demography and vital statistics that is considerably older than the modern mathematical discipline of statistics. Also old is the thought we might learn something about observational studies by their limited analogy with blocked or stratified or paired experiments. For instance, that thought is thematic in the

influential work of Nathan Mantel, who did not suggest it was a novel thought at the time he began writing in the 1950s.[8] In the 1960s and 1970s, William Cochran and Donald Rubin began the systematic development of observational studies as an organized aspect of statistical methodology, calling attention to both their similarities and dissimilarities with randomized experiments.[9] Rubin and I introduced propensity scores in the 1980s and developed their connection to ignorable treatment assignment.[10] Cochran's papers discuss elaborate theories, the topic of Chapter 7.

It would be difficult to trace the origin of natural experiments, the topic of Chapter 6, and every scientific field has its favorites. Natural experiments are used in epidemiology, economics, psychology and political science.[11] Many commonly used quasi-experimental devices were developed by Donald T. Campbell and colleagues, including multiple control groups, non-equivalent controls or counterparts, multiple time-series and regression-discontinuity designs.[12] Quasi-experimental techniques are widely used in public program evaluation, economics, finance and public health.[13] For simplicity in Chapter 8, I discussed quasi-experimental devices one-at-a-time, but studies often employ several devices in mutually supporting roles.[14] Evidence factors are a recent development.[15]

Sensitivity analysis in observational studies begins with the paper by Cornfield and colleagues, which is still worth reading. After 50 years, in 2009, it was republished in the *International Journal of Epidemiology* with four helpful current comments.[16] As indicated in the notes to Chapter 9, there are many recent proposals for sensitivity analysis in observational studies, and Chapter 9 has focused on one of these. Software for sensitivity analysis is widely available.[17] Design sensitivity is a comparatively recent development, but several surveys and examples are available.[18] Noel Weiss offers an interesting discussion of doses in observational studies.[19]

There are several reviews of matching methods.[20] Additionally, software for matching is widely available.[21] Generic biases and isolation in Chapter 12 are recent developments.[22]

In 1996, an influential paper by Joshua Angrist, Guido Imbens and Donald Rubin connected instruments with treatment effects in randomized experiments.[23] The discussion of effect ratios and the neonatal intensive care units example in Chapter 13 are based on a paper by Michael Baiocchi and colleagues.[24]

NOTES

Preface

1. Ronald A. Fisher, *Design of Experiments* (Edinburgh: Oliver and Boyd, 1935). The 1935 first edition is difficult to obtain, so when page numbers are given, they refer to the readily available fifth edition of 1949. For an appreciation of Frederick Mosteller, see Ivan Oransky, "Obituary: Charles Frederick Mosteller," *Lancet* 368 (2006): 1062.

2. For a technical discussion of causal inference in observational studies, see Paul R. Rosenbaum, *Observational Studies,* 2nd ed. (New York: Springer, 2002), and Paul R. Rosenbaum, *Design of Observational Studies* (New York: Springer, 2010).

3. John Stuart Mill, *A System of Logic* (1843; reprinted in Mill, *The Collected Works of John Stuart Mill,* vol. 7: *A System of Logic Ratiocinative and Inductive, Being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation* [books 1–3], ed. John M. Robson, intro. R. F. McRae [Toronto: University of Toronto Press; London: Routledge and Kegan Paul, 1974]). In two distinct roles, Mill makes cameo appearances in *Observation and Experiment* (see, for instance, Chapter 10). In his first role, Mill is the author of *A System of Logic,* a book of historical importance in which, with his characteristic precision and force, he articulated a view of causal inference that is highly intuitive and common, yet mistaken, which involves comparing identical people under competing treatments. In his second role, Mill is the author of *On Liberty,* a living book in which, with

his characteristic precision and force, he argued that knowledge worthy of the name is possible only if purported knowledge is subjected to critical challenge.

4. O. Ashenfelter and C. Rouse, "Income, schooling and ability: Evidence from a new sample of identical twins," *Quarterly Journal of Economics* 113 (1998): 253–284; Avshalom Caspi, Julia Kim-Cohen, Terrie E. Moffitt, Julia Morgan, Michael Rutter, Monica Polo-Tomas, and Alan Taylor, "Maternal expressed emotion predicts children's antisocial behavior problems: Using monozygotic-twin differences to identify environmental effects on behavioral development," *Developmental Psychology* 40 (2004): 149–161; A. Haapanen, M. Koskenvuo, J. Kaprio, Y. A. Kesäniemi, and K. Heikkilä, "Carotid arteriosclerosis in identical twins discordant for cigarette smoking," *Circulation* 80 (1989): 10–16.

5. Fisher, *Design of Experiments,* 18.

6. Paul Meier, "The biggest public health experiment ever: The 1954 field trial of the Salk vaccine," in *Statistics: A Guide to the Unknown,* ed. Judith M. Tanur, 2–13 (San Francisco: Holden-Day, 1972).

7. Richard Doll and Austin Bradford Hill, "The mortality of doctors in relation to their smoking habits," *British Medical Journal* 1 (1954): 1451–1455.

8. Austin Bradford Hill, "Observation and experiment," *New England Journal of Medicine* 248 (1953): 3–9.

9. U.S. Public Health Service, *Smoking and Health: Report of the Advisory Committee to the Surgeon General* (Washington, DC: Department of Health, Education and Welfare, 1964).

10. William G. Cochran, "The planning of observational studies of human populations (with discussion)," *Journal of the Royal Statistical Society,* Series A, 128 (1965): 234–266.

11. Italo Calvino, *Six Memos for the Next Millennium* (Cambridge, MA: Harvard University Press, 1988), 3, 16.

Reading Options

1. For instance, random assignment of I people to either treatment or control by flipping a fair coin is slightly different from randomly picking m of I people for treatment, assigning the remaining $I - m$ people to control. Conceptually, this distinction is a minor matter, but it does have some consequences for technical details that are unlikely to interest most readers. In the text of the book, I pay little attention to this distinction, but I clarify the distinction in a few endnotes. We may pass from the first situation to the second by fixing m by conditioning on its observed value, and this is almost invariably done in practice. So my discussion of this minor matter is consistent with my policy of using endnotes to discuss conditional probabilities.

2. "One accepts one's destiny—that's what destiny is. Justino knows and accepts this. Otherwise he would be able to fight it." Pedro Rosa Mendes, *Bay of Tigers* (New York: Harcourt, 2003), 13.

3. This is the potential outcomes framework for causal effects introduced by Jerzy Neyman and Donald Rubin. See Chapters 1 and 2.

Book Epigraphs

1. John Dewey, *Essays in Experimental Logic* (Chicago: University of Chicago Press, 1916).

2. Alva Noë, *Varieties of Presence* (Cambridge, MA: Harvard University Press, 2012), 2.

3. Robert P. Crease, *The Prism and the Pendulum: The Ten Most Beautiful Experiments in Science* (New York: Random House, 2004), xv.

1. A Randomized Trial

1. Emanuel Rivers, Bryant Nguyen, Suzanne Havstad, Julie Ressler, Alexandria Muzzin, Bernhard Knoblich, Edward Peterson, and Michael Tomlanovich for the Early Goal-Directed Therapy Group, "Early goal-directed therapy in the treatment of severe sepsis and septic shock," *New England Journal of Medicine* 345 (2001): 1368–1377.

2. ProCESS Investigators, "A randomized trial of protocol-based care for early septic shock," *New England Journal of Medicine* 370 (2014): 1683–1693. Although the example in this chapter is based on this reference, it simplifies certain features of the original experiment. The experiment had three treatment groups, but only two are discussed here. In particular, the "usual care" treatment group is not discussed. Discussing three treatments rather than two would add detail that is important to critical care medicine, but it would not introduce new aspects of causal inference. The trial used a centrally administered stratified block randomization rather than actually flipping coins, but I will ignore this technical subtlety. At various moments, I will ignore a second related subtly, namely, the distinction between assigning treatments through independent coin flips or assigning a random half of the patients to one treatment, the other half to the alternative treatment.

3. Sir Ronald A. Fisher, *Design of Experiments* (Edinburgh: Oliver and Boyd, 1935).

4. Of course, there could be three treatments, or seven treatments, or a continuum of treatments at different doses, or regimens of treatment extending over time, but

these would require minor adjustments without new ideas. The case of two treatments is the simplest nontrivial case.

5. In statistics, the idea of defining causal effects as comparisons of potential outcomes under alternative treatments is typically credited to Jerzy Neyman and Donald Rubin. Neyman introduced the idea in the context of randomized experiments in the 1920s, and Rubin developed the idea in other areas of causal inference. See, for example, Jerzy Neyman, "On the application of probability theory to agricultural experiments: Essay on principles" (in Polish), *Roczniki Nauk Roiniczych Tom* 10 (1923): 1–51, reprinted in English in *Statistical Science* 5 (1990): 463–480; B. L. Welch, "On the z-test in randomized blocks and Latin squares," *Biometrika* 42 (1937): 21–52; and Donald B. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies," *Journal of Educational Psychology* 66 (1974): 688–701.

6. Paul R. Rosenbaum, "The consequences of adjustment for a concomitant variable that has been affected by the treatment," *Journal of the Royal Statistical Society,* Series A, 147 (1984): 656–666.

7. W. A. Knaus, D. P. Wagner, and J. Lynn, "Short-term mortality predictions for critically ill hospitalized patients," *Science* 254 (1991): 389–394; D. P. Wagner, W. A. Knaus, F. E. Harrell, J. E. Zimmerman, and C. Watts, "Daily prognostic estimates for critically ill adults in intensive care units," *Critical Care Medicine* 22 (1994): 1359–1372.

8. In the ProCESS Trial, the primary outcome was the rate of in-hospital death from any cause at 60 days. A secondary outcome was the rate of death from any cause at 90 days. Because these two outcomes yielded similar conclusions, I will not emphasize the distinction between them in this book. Each outcome had a small flaw, but because they agreed, we are not greatly worried by their small flaws in this particular experiment. A handful of people were not recorded in the mortality rate at 90 days, perhaps because they were lost to follow-up evaluation after discharge from the hospital. Hospitals always know whether a patient has died in the hospital, so in-hospital mortality rates do not suffer from missing data, but they are slightly peculiar as mortality rates. The in-hospital death rate from any cause at 60 days contrasts two groups: people who died in the hospital before 60 days and people who were either discharged from the hospital alive before 60 days or were alive in the hospital at 60 days.

9. Neyman, "On the application of probability theory"; Welch, "On the z-test"; Rubin, "Estimating causal effects."

10. John Stuart Mill, *A System of Logic* (1843; reprinted in Mill, *The Collected Works of John Stuart Mill,* vol. 7: *A System of Logic Ratiocinative and Inductive, Being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation* [books 1–3], ed. John M. Robson, intro. R. F. McRae [Toronto: University of Toronto Press; London: Routledge and Kegan Paul, 1974]).

11. Rosenbaum, "Consequences of adjustment."

12. One sense in which luck can be reliable is given by a theorem—actually a collection of related theorems—called the "law of large numbers." A simple version of the law of large numbers applies to flipping the same coin independently many times. In this case, the law of large numbers says the proportion of heads tends to the probability of a head as the number of flips increases. For an introduction to the law of large numbers, see volume 1 of William Feller, *An Introduction to Probability and Its Applications* (New York: John Wiley, 1968).

13. Forcing balance while also randomizing is easy to do if everyone enters the trial at the same time. For instance, if you wanted to force balance for gender, you could randomly pick half the men for treatment, then randomly pick half the women for treatment. Then the number of treated men and the number of control men would be equal if the number of men was even, or it would differ by one if the number of men was odd. The same would be true for women. Hence, the proportion of men in the treated group would be almost exactly the same as the proportion of men in the control group. Notice that if you balance the number of men and balance the number of women, you also balance the proportion of men. This is known as "blocked randomization," as distinct from the "complete randomization" in the ProCESS Trial; see David R. Cox, *Planning of Experiments* (New York: John Wiley and Sons, 1958), chapter 3.

What if you wanted to force balance on many covariates at the same time while also randomizing the treatment assignment? This is possible by pairing people for many covariates and then flipping a coin to assign one person in each pair to treatment, the other to control; see Robert Greevy, Bo Lu, Jeffrey H. Silber, and Paul R. Rosenbaum, "Optimal multivariate matching before randomization," *Biostatistics* 5 (2004): 263–275.

It is somewhat more difficult to force balance when people enter the trial gradually over a long period of time, as was true in the ProCESS Trial, because you do not know how many men and how many women will ultimately enter the trial. In this case, Bradley Efron suggested flipping a biased coin, not a fair coin, where the bias is tilted to correct the current imbalance in covariates; see Bradley Efron, "Forcing a sequential experiment to be balanced," *Biometrika* 58 (1971): 403–417.

14. Fisher, *Design of Experiments,* chapter 2.

15. An alert reader will notice that something has just slipped. Up until now, we have spoken of assigning treatments by the flip of a fair coin. A fair coin might put 439 people in the aggressive treatment group, as it did in the ProCESS Trial. If we imagine the many such experiments we might have gotten with different flips of our fair coin, some would have 439 people in the aggressive treatment group, but others would put a slightly different number of patients in the aggressive treatment group. A different kind of randomized trial—more like a lottery with a fixed number of winning tickets—would pick 439 patients at random from the 885 patients so that no matter how many times the lottery is run, there are always 439 equitably chosen winners. In

this different kind of randomized experiment, every group of 439 patients has the same chance of being chosen for aggressive treatment.

When we analyze data from a coin-flip experiment, we usually act as if it was a fair lottery instead. This "act as if" is actually conditioning on the event that 439 patients were assigned to the aggressive treatment, a concept from probability theory. This topic is subtle and not critical to an understanding of causal inference, so I will not discuss it in this book—after all, the fair lottery experiment can be performed, and often is performed. The ProCESS Trial did not hold a fair lottery because patients entered the trial gradually over a long period of time; to hold a fair lottery, you need to decide the winning tickets all at once.

2. Structure

1. George Polya, *How To Solve It* (Princeton, NJ: Princeton University Press, 1957), 134.

2. The sense in which u_i may record more than one covariate is developed in the endnotes to later chapters, particularly Chapter 9.

3. Jerzy Neyman, "On the application of probability theory to agricultural experiments: Essay on principles" (in Polish), *Roczniki Nauk Roiniczych Tom* 10 (1923): 1–51, reprinted in English in *Statistical Science* 5 (1990): 463–480; Donald B. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies," *Journal of Educational Psychology* 66 (1974): 688–701.

4. The phrase "under mild conditions" often appears in mathematical or statistical claims. It means that some general statement is indeed true under very general circumstances but not under all circumstances, and the speaker or writer does not want to get into the details at the moment. What does the phrase "under mild conditions" refer to here? When is the mean of a random sample *not* a good estimate of the mean of a population? If, somehow, we could draw up a list of all the objects in our solar system and draw a random sample of half of them, then the average mass of the objects in our sample would be a very poor estimate of the average mass in the population. The reason is that our population includes one object, the sun, that is so extreme in its mass; it is much more massive than every other object in our population. If our random sample of half the population excluded the sun, then the sample mean would be way too low as an estimate of the population mean. If our sample included the sun, then the sample mean would be way too high as an estimate of the population mean, essentially because the sample mean would think there is another sun in the half of the population we did not see. I have given an extreme example, but in a country with substantial income inequality, similar but less extreme things happen when you estimate the population mean income of citizens of a country using a random sample: missing a few people can have a big impact on a mean.

5. William G. Cochran, *Sampling Techniques* (New York: John Wiley and Sons, 1963).

6. Strictly speaking, we should not write $\hat{r}_T$ for $(1/m)(Z_1\, r_{T1} + \cdots + Z_I\, r_{TI})$ because the T appears once but signifies two things, namely averaging treated responses over the treated group. This notation could create confusion in some contexts that do not arise in this book.

7. The hypothesis, H_0: $\bar{\delta} = 0$, that the average treatment effect is zero, is a composite null hypothesis; that is, there are many possible δ_i that make it true. Fisher's hypothesis, H_0: $\delta_i = 0$ for all i, is one component of this composite; it is one way that H_0: $\bar{\delta} = 0$ can be true. To reject a composite hypothesis is to reject each and every way it may be true—that is, to reject each component of the composite. So to reject H_0: $\bar{\delta} = 0$ you must reject Fisher' hypothesis. However, it turns out that in a paired randomized experiment Fisher's hypothesis is the component that is most difficult to reject, so having rejected Fisher's hypothesis you have rejected H_0: $\bar{\delta} = 0$. The previous statement is a bit technical, and there are some details in the fine print. For a result of this form in the randomized paired case, see proposition 2 in Mike Baiocchi, Dylan S. Small, Scott Lorch, and Paul R. Rosenbaum, "Building a stronger instrument in an observational study of perinatal care for premature infants," *Journal of the American Statistical Association* 105 (2010): 1285–1296. Their result is more general than inference about average treatment effects; it covers also effect ratios and instrumental variables as discussed in Chapter 13. The case of an average treatment effect is the case of an instrument with perfect compliance.

3. Causal Inference in Randomized Experiments

1. William G. Cochran, "Catalogue of uniformity trial data," *Journal of the Royal Statistical Society,* Suppl. 4 (1937): 233–253.

2. As I have mentioned a few times in footnotes, in randomization inference, we usually act as if m were fixed, so we pick $m = 4$ people at random from $I = 8$ people, and there are 70 ways to make the split. We could do this with coin flips by flipping a fair coin independently until four people are assigned to either treatment or control, assigning the rest to the group that has not yet received four people. If we did not insist that there are $m = 4$ people of $I = 8$ people in the treated group, there would be $2^I = 2^8 = 256$ ways to assign treatments to $I = 8$ people, including some silly ways such as assigning everyone to the control group. If we used the second method, with independent coin flips and 256 possible assignments, but we happened to get $m = 4$ people in the treated group, we could still legitimately act as if we had randomized a fixed number, $m = 4$, of people to receive treatment by conditioning on the observed value of the random variable m. It is an elementary exercise in probability to show that, with independent coin flips, the conditional distribution of treatment assignments

given $m = 4$ is the same as the randomization distribution that arises when we pick a fixed number, $m = 4$, to receive treatment. It is for this reason that we do not worry much about whether m is fixed by design or fixed by conditioning on its observed value: in a randomized experiment, we obtain the same distribution of treatment assignments either way.

3. An attractive, conceptual survey of significance levels and P-values is given by David R. Cox, "The role of significance tests (with discussion)," *Scandinavian Journal of Statistics* 4 (1977): 49–70. I follow Cox in defining the two-sided P-value as twice the smaller of two one sided P-values or 1, whichever is smaller, essentially testing two one-sided hypotheses and correcting for testing two hypotheses using the Bonferroni inequality. There are subtle aspects of two-sided P-values for asymmetric null distributions that I will not discuss in this book; see Juliet P. Shaffer, "Bidirectional unbiased procedures," *Journal of the American Statistical Association* 69 (1974): 437–439. The standard reference for hypothesis testing is Erich L. Lehmann and Joseph Romano, *Testing Statistical Hypotheses,* 3rd ed. (New York: Springer, 2005). That book contains in its fifth chapter a mathematically beautiful, if slightly technical, discussion of randomization inference and permutation tests.

4. Larry V. Hedges and Ingram Olkin, *Statistical Methods for Meta-Analysis* (New York: Academic Press, 1985); Matthias Egger, George Davey-Smith, and Douglas Altman, eds., *Systematic Reviews in Health Care: Meta-Analysis in Context* (London: BMJ, 2001).

5. The term "false discovery" is from Yoav Benjamini and Yosef Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society,* Series B, 57 (1995): 289–300. They proposed a new way to control the proportion of discoveries that were false when many hypotheses were tested at the same time.

6. For a textbook discussion of equivalence testing, see Stephan Welleck, *Testing Hypotheses of Equivalence and Noninferiority,* 2nd ed. (Boca Raton, FL: Chapman and Hall / CRC, 2010). For a survey paper with discussion, see Roger L. Berger and Jason C. Hsu, "Bioequivalence trials, intersection-union tests and equivalence confidence sets," *Statistical Science* 11 (1996): 283–319. For a concise but technical discussion, see P. Bauer and M. Kieser, "A unifying approach for confidence intervals and testing of equivalence and difference," *Biometrika* 83 (1996): 934–937. For a simple but extremely clever way to coordinate tests for effect and for equivalence, see Jelle J. Goeman, Aldo Solari, and Theo Stijnen, "Three-sided hypothesis testing: Simultaneous testing of superiority, equivalence and inferiority," *Statistics in Medicine* 29 (2010): 2117–2125. For an illustration of the use of three-sided testing with an attributable effect, see Samuel D. Pimentel, Rachel R. Kelz, Jeffrey H. Silber, and Paul R. Rosenbaum, "Large, sparse optimal matching with refined covariate balance in an observational study of the health outcomes produced by new surgeons," *Journal of the American Statistical Association* 110 (2015): 515–527.

7. Confidence intervals were proposed and built from hypothesis tests by Jerzy Neyman. For discussion of the construction of confidence intervals from hypothesis tests—the so-called duality of tests and confidence intervals—see Lehmann and Romano, *Testing Statistical Hypotheses,* chapter 3.

8. In technical terms, we say that the nominal level of the test is 0.05, while the actual size of the test is 0.0143, and a test is a valid test if its actual size is at most equal to its nominal level. The reader concerned with the distinction between size and level will find extensive discussion in Lehmann and Romano, *Testing Statistical Hypotheses,* chapter 3.

9. David R. Cox (1966), "A simple example of a comparison involving quantal data," *Biometrika* 52 (1966): 213–220.

10. For a related discussion, see John W. Tukey, "Sunset salvo," *American Statistician* 40 (1986): 72–76.

11. As noted previously, this is not exactly the way the randomization was done in the ProCESS Trial, because patients entered the ProCESS Trial gradually over time and could not be randomized at a single moment. I am ignoring some minor technical subtleties that separate the small example from the ProCESS Trial.

12. See David R. Cox, *The Analysis of Binary Data* (London: Methuen, 1970). The statistical software package R is freely available from https://cran.r-project.org. In R, one obtains the two-sided *P*-value of 0.3352 as:

```
> ProCESS<-matrix(c(92,81,347,365),2,2)
> fisher.test(ProCESS,alternative="greater")
        Fisher's Exact Test for Count Data
data:   ProCESS
p-value = 0.1676
> 2*0.1676
[1] 0.3352
```

13. There are many methods for estimating the magnitude of an effect in a 2×2 table. Most such methods introduce additional structures and sources of randomness that are not part of Fisher's randomization inference, such as samples from infinite populations and probability models; see, for instance, Joseph L. Fleiss, Bruce Levin, and Myunghee Cho Paik, *Statistical Methods for Rates and Proportions* (New York: John Wiley and Sons, 2013). The method in the current section retains the flavor of Fisher's discussion in that probability enters only through the random assignment of treatments. The method is described in Paul R. Rosenbaum, "Effects attributable to treatment: Inference in experiments and observational studies with a discrete pivot," *Biometrika* 88 (2001): 219–231. The method illustrates the elementary fact that we may make confidence statements about the magnitude of a treatment effect based solely on the random assignment of treatments, without introducing a model for the responses. The method is easy to use, but if you plan to use it, I suggest reading the article first because there is more to the story than is presented in Chapter 3. Extensions

of this method are discussed in Paul R. Rosenbaum, "Attributing effects to treatment in matched observational studies," *Journal of the American Statistical Association* 457 (2002): 183–192; and Colin B. Fogarty, Pixu Shi, Mark E. Mikkelsen, and Dylan S. Small, "Randomization inference and sensitivity analysis for composite null hypotheses with binary outcomes in matched observational studies," *Journal of the American Statistical Association* 112 (2017), forthcoming. A general discussion of measures of the magnitude of causal effects in 2×2 tables is given by Martin A. Hamilton, "Choosing the parameter for a 2×2 or a 2×2×2 table analysis," *American Journal of Epidemiology* 109 (1979): 362–375.

14. It is for this same reason that we do not correct for testing many hypotheses when constructing a 95% confidence interval for a parameter θ: we test every value of θ, but we test a true value only once.

15. In R we compute:

```
> ProCESS<-matrix(c(92+13,81,347−13,365),2,2)
> fisher.test(ProCESS,alternative="greater")
p−value = 0.02167
> 2*0.02167
[1] 0.04334
```

16. In R we compute:

```
> ProCESS<-matrix(c(92−35,81,347+35,365),2,2)
> fisher.test(ProCESS,alternative="less")
p−value = 0.02101
> 2*0.02101
[1] 0.04202
```

17. The calculation just described yields a two-sided 95% confidence interval for the attributable effect *A*. The process of building a confidence interval by testing hypotheses is called "inverting the test," and it defines the relationship between tests and confidence intervals; see Lehmann and Romano, *Testing Statistical Hypotheses,* chapter 3. The quantity *A* is an unknown random variable, but confidence intervals for random variables are similar to confidence intervals for fixed parameters; see Lionel Weiss, "A note on confidence sets for random variables," *Annals of Mathematical Statistics* 26 (1955): 142–144. If a two-sided confidence interval is replaced by a two-sided test for difference and a test for equivalence, a narrower confidence set for equivalence is obtained using the method of Goeman et al., "Three-sided hypothesis testing."

4. Irrationality and Polio

1. Daniel Kahneman and Amos Tversky, "Choices, frames and values," *American Psychologist* 39 (1984): 341–350; Amos Tversky and Daniel Kahneman, "Rational choice and the framing of decisions," *Journal of Business* 59 (1986): S251–S278.

2. Barbara J. McNeil, Stephen G. Pauker, Harold C. Sox Jr., and Amos Tversky, "On the elicitation of preferences for alternative therapies," *New England Journal of Medicine* 306 (1982): 1259–1262.

3. McNeil et al., "On the elicitation of preferences," 1260; Tversky and Kahneman, "Rational choice," S254.

4. The discussion and numerical results of the Salk vaccine trial draw heavily from an essay by Paul Meier, "The biggest public health experiment ever: The 1954 field trial of the Salk vaccine," in *Statistics: A Guide to the Unknown,* ed. Judith M. Tanur et al., 2–13 (San Francisco: Holden-Day, 1972). See also the following references: Thomas Francis Jr., et al. "Evaluation of the 1954 poliomyelitis vaccine trials," *American Journal of Public Health* 45 (1955): 1–51; K. Alexander Brownlee, "Statistics of the 1954 polio vaccine trials," *Journal of the American Statistical Association* 50 (1955): 1005–1013; Paul Meier, "Polio trial: An early efficient clinical trial," *Statistics in Medicine* 9 (1990): 13–16; Marcia Meldrum, "A calculated risk: The Salk polio vaccine field trials of 1954," *British Medical Journal* 317 (1998): 1233–1236.

5. Lincoln Moses, "Measuring effects without randomized trials? Options, problems, challenges," *Medical Care* 33 (1995): AS8–AS14, quotation from p. A12.

6. Francis et al., "Evaluation of the 1954 poliomyelitis vaccine trials," table 1a.

7. It may seem surprising, but experimenters often inadvertently distort the treatment effects they are studying. See Robert Rosenthal, *Experimenter Effects in Behavioral Research* (East Norwalk, CT: Appleton-Century-Crofts, 1966).

8. Leila Calhoun Deasy, "Socio-economic status and participation in the poliomyelitis vaccine trial," *American Sociological Review* 21 (1956): 185–191, see table 1.

9. Deasy, "Socio-economic status," tables 3 and 6.

10. Francis, "Evaluation of the 1954 poliomyelitis vaccine trials," 5.

11. The *P*-value testing the equality of polio rates in the trial and the observational study is 0.072. This is based on a one-degree of freedom likelihood ratio test in the $2 \times 2 \times 2$ table recording paralytic polio, vaccination and trial or observational study, comparing conditional independence with constant partial association. In parallel, the P-value testing equality of odds ratios in the trial and observational study is 0.304, obtained by comparing constant partial association to the saturated model.

12. Meier, "Biggest public health experiment," table 1; Francis, "Evaluation of the 1954 poliomyelitis vaccine trials," table 2b.

13. Meier, "Biggest public health experiment," 8.

5. Between Observational Studies and Experiments

1. John Dewey, *Reconstruction in Philosophy* (New York: Dover, 2004), 7.

2. William G. Cochran, "The planning of observational studies of human populations (with discussion)," *Journal of the Royal Statistical Society,* Series A, 128 (1965): 134–155.

3. E. H. Simpson, "The interpretation of interaction in contingency tables," *Journal of the Royal Statistical Society,* Series B, 13 (1951): 238–241. For some related discussion, see P. J. Bickel, E. A. Hammel, and J. W. O'Connell, "Sex bias in graduate admissions: Data from Berkeley," *Science* 187 (1975): 398–404.

4. Direct adjustment of rates is an old topic. For a modern discussion, including standard errors and methods of inference, see Joseph L. Fleiss, Bruce Levin, and Myunghee Cho Paik, *Statistical Methods for Rates and Proportions* (New York: John Wiley and Sons, 2013).

5. It is not difficult to demonstrate this with some routine algebra; see, for instance, section 2.7.1 of my *Observational Studies,* 2nd ed. (New York: Springer, 2002). Specifically, it is shown there that direct adjustment with stratum total weights yields a conditionally unbiased estimate of the average treatment effect if treatments are randomly assigned with probabilities that are constant within each stratum but vary from one stratum to another.

6. When might one use weights that are not proportional to the stratum size? One might be interested in the effect of the treatment on the people who received treatment. In this case, the weights would be proportional to the number of treated people in a stratum. In Table 5.1, this means that stratum 1, older men, would receive weight $0.4003 = 79,828 / (79,828 + 79,779 + 20,021 + 19,798)$, because older men constitute roughly 40% of the people who received treatment, though only 25% of the overall population. On the other hand, if you were thinking of expanding use of the treatment so that the controls received it too, then you might be interested in the average effect of the treatment on controls, and then stratum 1 would receive weight $0.1006 = 20,172 / (20,172 + 20,221 + 79,979 + 80,202)$. In Table 5.1, there is no treatment effect, so direct adjustment estimates zero effect for every set of weights. If treatments were being given to people in a sensible way, as they might be in some well-established medical practice, then people who currently receive treatment might tend to benefit more from treatment than people who currently receive control, so the effect of the treatment on treated patients might be very different from the effect of treatment on controls.

7. The standard test for Table 5.1 is the Mantel-Haenszel test, introduced in: N. Mantel and W. Haenszel, "Statistical aspects of retrospective studies of disease," *Journal of the National Cancer Institute* 22 (1957): 719–748. The test uses a randomization distribution in each stratum, in parallel with Chapter 3, but it does not assume the same treatment assignment probabilities in different strata. For enlightening discussions of the Mantel-Haenszel test and related matters, see M. W. Birch, "The detection of partial association, I: The 2×2 case," *Journal of the Royal Statistical Society,* Series B, 26 (1964): 313–324; and David R. Cox, "A simple example of a comparison involving quantal data," *Biometrika* 53 (1966): 215–220. The discussion by Cox is particularly clear about the connection with randomization distributions. The reasoning here is a very special case of a general argument given in Paul R. Rosenbaum,

"Conditional permutation tests and the propensity score in observational studies," *Journal of the American Statistical Association* 79 (1984): 565–574.

8. Indeed, even if we knew the treatment assignment probabilities, there are good reasons to ignore them and instead base inferences solely on the fact that each stratum of Table 5.1 is a separate completely randomized experiment. This issue is discussed in greater detail in Rosenbaum, "Conditional permutation tests," and in Paul R. Rosenbaum, "Model-based direct adjustment," *Journal of the American Statistical Association* 82 (1987): 387–394. Expressed in slightly technical terms, it is important to know that treatment assignment is ignorable given the strata, but not important to know the value of the propensity score within strata.

9. Donald B. Rubin, "Assignment to treatment group on the basis of a covariate," *Journal of Educational Statistics* 2 (1977): 1–26.

10. Rosenbaum, *Observational Studies,* chapter 3.

11. T. Gordon and B. M. Foss, "The role of stimulation in the delay of onset of crying in the newborn infant," *Quarterly Journal of Experimental Psychology* 18 (1966): 79–81. The experiment was discussed by John Bowlby in his study of emotional development in small children, *Attachment and Loss,* vol. 1 (New York: Basic Books, 1982), 293, and by Cox, "Simple example of a comparison." In particular, Cox uses essentially the Mantel-Haenszel test in the analysis of the data.

12. Gordon and Foss, "The role of stimulation," 80.

13. Interference between units was discussed by David R. Cox, *Planning of Experiments* (New York: John Wiley and Sons, 1958) and Donald B. Rubin, "Which ifs have causal answers," *Journal of the American Statistical Association* 81 (1986): 961–962. Causal inference in the presence of interference is discussed by Michael E. Sobel, "What do randomized studies of housing mobility demonstrate? Causal inference in the face of interference," *Journal of the American Statistical Association* 101 (2006): 1398–1407; Paul R. Rosenbaum, "Interference between units in randomized experiments," *Journal of the American Statistical Association* 102 (2007): 191–200; Michael Hudgens and M. Elizabeth Halloran, "Toward causal inference with interference," *Journal of the American Statistical Association* 103 (2012): 832–842; Xi Luo, Dylan S. Small, Chiang-Shan R. Li, and Paul R. Rosenbaum, "Inference with interference between units in an fMRI experiment of motor inhibition," *Journal of the American Statistical Association* 103 (2012): 530–541; Jake Bowers, Mark M. Fredrickson, and Costas Panagopoulos, "Reasoning about interference between units," *Political Analysis* 21 (2013): 97–124; Lan Liu and Michael G. Hudgens, "Large sample randomization inference for causal effects in the presence of interference, *Journal of the American Statistical Association* 109 (2014): 288–301; and David Choi, "Estimation of monotone treatment effects in network experiments," *Journal of the American Statistical Association* (2016) doi: 10.1080/01621459.2016.1194845. In particular, Luo et al., "Inference with interference," analyze a massive randomized experiment similar to the crying babies experiment, in which a small number of people were randomized

many times to cognitive stimuli while their brains were monitored by functional magnetic resonance imaging (fMRI).

14. William G. Cochran, *Sampling Techniques* (New York: John Wiley and Sons, 1963).

15. National Center for Education Statistics, U.S. Department of Education, *High School and Beyond,* http://nces.ed.gov/surveys/hsb/.

16. See, for instance, Eugene S. Edgington, "Randomized single-subject experiments and statistical tests," *Journal of Counseling Psychology* 34 (1987): 437–442.

17. Allan Donner and Neil Klar, *Design and Analysis of Cluster Randomization Trials in Health Research* (New York: John Wiley and Sons, 2010).

18. See Martha L. Bruce, Thomas R. Ten Have, Charles F. Reynolds III, Ira I. Katz, Herbert C. Schulberg, Benoit H. Mulsant, Gregory K. Brown, Gail J. McAvay, Jane L. Pearson, and George S. Alexopoulos, "Reducing suicidal ideation and depressive symptoms in depressed older primary care patients: A randomized controlled trial," *Journal of the American Medical Association* 291 (2004): 1081–1091; and Dylan S. Small, Thomas R. Ten Have, and Paul R. Rosenbaum, "Randomization inference in a group–randomized trial of treatments for depression: Covariate adjustment, noncompliance, and quantile effects," *Journal of the American Statistical Association* 103 (2008): 271–279.

19. Less intuitively but more precisely, the conditional probability that Harry received treatment given that exactly one of Harry and Sally received treatment is $1/2 = \Pr(Z_1 = 1 \text{ and } Z_2 = 0 \mid Z_1 + Z_2 = 1)$. Why is this true? Well, $\Pr(Z_1 = 1 \text{ and } Z_2 = 0) = \pi_1(1 - \pi_2)$ by independence of the two coin flips, and $\Pr(Z_1 = 0 \text{ and } Z_2 = 1) = \pi_2(1 - \pi_1)$ for the same reason. Also, $\Pr(Z_1 + Z_2 = 1) = \Pr(Z_1 = 1 \text{ and } Z_2 = 0) + \Pr(Z_1 = 0 \text{ and } Z_2 = 1)$ because these are the two mutually exclusive ways to have exactly one treated person among the first two people. So, we have $\Pr(Z_1 + Z_2 = 1) = \pi_1(1 - \pi_2) + \pi_2(1 - \pi_1)$. By the definition of conditional probability, it follows that $\Pr(Z_1 = 1 \text{ and } Z_2 = 0 \mid Z_1 + Z_2 = 1)$ is

$$\frac{Pr(Z_1 = 1 \text{ and } Z_2 = 0)}{Pr(Z_1 + Z_2 = 1)} = \frac{\pi_1(1 - \pi_2)}{\{\pi_1(1 - \pi_2) + \pi_2(1 - \pi_1)\}},$$

which equals $1/2$ whenever $\pi_1 = \pi_2$.

20. Sharon K. Inouye, Sidney T. Bogardus Jr., Peter A. Charpentier, Linda Leo-Summers, Denise Acampora, Theodore R. Holford, and Leo M. Cooney Jr, "A multicomponent intervention to prevent delirium in hospitalized older patients," *New England Journal of Medicine* 340 (1999): 669–676.

21. Strictly speaking, the treatment in this study is treatment in the unit in which the Elder Life Program was available, and the control was treatment in one of the units in which it was not available. It is conceivable, at least in principle, that it was

not the Elder Life Program but some other difference in care among the programs that is responsible for any treatment effect.

22. In the free statistical software package R:

```
> pbinom(33,88,.5)
[1] 0.01231142
> 2*pbinom(33,88,.5)
[1] 0.02462283
```

23. See David R. Cox, *Analysis of Binary Data* (London: Methuen, 1970).

24. Unlike Table 5.5, the $2 \times 2 \times P$ contingency table counts people, not pairs. The rows are treatment, $Z_i = 1$, or control, $Z_i = 0$, the columns record the binary outcome, and the strata are the pairs. Each row contains one person because each pair contains one treated person and one control. A pair is concordant if one column of the table contains two people and the other column contains no people, and it is discordant if each column contains one person. The Mantel-Haenszel test applied to this $2 \times 2 \times P$ contingency is equivalent to McNemar's test.

25. Propensity scores were introduced in terms of samples from infinite populations in Paul R. Rosenbaum and Donald B. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika* 70 (1983): 41–55. The discussion of propensity scores in this book keeps a closer link with Fisher's theory of randomized experiments by viewing propensity scores in terms of the finite population of I individuals in the current study, and in that sense the discussion in this book follows that in Rosenbaum, "Conditional permutation tests," and Rosenbaum, *Observational Studies,* chapter 10. The main properties of the propensity score are the same in either formulation.

26. For a proof of the balancing property of propensity scores in infinite populations, see Rosenbaum and Rubin, "Central role of the propensity score," theorem 1, whereas for a proof of the property in finite populations, see Rosenbaum, *Observational Studies,* chapter 10, proposition 29. The proof is a little too technical for this book, but it is just a few lines long and not particularly technical.

27. This phenomenon is often observed when propensity scores are used, but it has a theoretical explanation. It is a well-known fact in survey sampling that if you pretend that a simple random sample is a stratified random sample, a process called post-stratification, then you often produce a more stable estimate of the population mean than is provided by the sample mean; see D. Holt and T. M. F. Smith, "Post stratification," *Journal of the Royal Statistical Society,* Series A, 142 (1979): 33–46. Post-stratification pretends stratum-specific sampling probabilities equal their sample proportions: it replaces known sampling probabilities by estimates, yet the resulting estimator of the population mean is often more, rather than less, precise. The same phenomenon occurs with estimated propensity scores; see Rosenbaum, "Model-based direct adjustment." Moreover, a parallel phenomenon occurs when testing the null

hypothesis of no treatment effect, namely, estimated propensity scores are slightly better than true propensity scores; see Rosenbaum, "Conditional permutation."

28. Paul R. Rosenbaum and Donald B. Rubin, "Reducing bias in observational studies using subclassification on the propensity score," *Journal of the American Statistical Association* 79 (1984): 516–524.

29. The model was a logit model, a standard model predicting a binary variable: here, treatment Z from predictors; here, the observed covariates x. For discussion of logit models, see D. R. Cox and E. J. Snell, *Analysis of Binary Data,* 2nd ed. (New York: Chapman and Hall, 1989).

30. Rosenbaum and Rubin, "Central role of the propensity score," theorem 2.

31. Paul R. Rosenbaum and Donald B. Rubin, "Constructing a control group using multivariate matched sampling methods that incorporate the propensity score," *American Statistician* 39 (1985): 33–38.

32. The stratification that contains both treated and control subjects and is as close as possible on many covariates is invariably a form of matching called "full matching," with one treated subject and one or more controls or one control and one or more treated subjects; see Paul R. Rosenbaum, "A characterization of optimal designs for observational studies," *Journal of the Royal Statistical Society,* Series B, 53 (1991): 597–610.

33. Here is a slightly more formal definition. Treatment assignment is defined to be ignorable given x if $0 < \Pr(Z_i = 1|x_i, r_{Ti}, r_{Ci}) = \Pr(Z_i = 1|x_i) < 1$ for $i = 1, \ldots, I$. Here, $\pi_i = \Pr(Z_i = 1|x_i, r_{Ti}, r_{Ci})$, and the propensity score is $\lambda(x_i) = \Pr(Z_i = 1|x_i)$. So we may restate this as: treatment assignment is ignorable given x if $0 < \pi_i = \lambda(x_i) < 1$ for $i = 1, \ldots, I$. We often express failures of ignorable treatment assignment in terms of an unobserved covariate, u_i, such that treatment assignment is ignorable given (x,u) but not given x. When treatment assignment is not ignorable, then there is always an unobserved covariate u_i with $0 \le u_i \le 1$ such that $0 \le \Pr(Z_i = 1|x_i, u_i, r_{Ti}, r_{Ci}) = \Pr(Z_i = 1|x_i, u_i) \le 1$ for $i = 1, \ldots, I$, namely the unobserved covariate $u_i = \pi_i = \Pr(Z_i = 1|x_i, r_{Ti}, r_{Ci})$. In other words, if $0 < \Pr(Z_i = 1|x_i, r_{Ti}, r_{Ci}) < 1$, for $i = 1, \ldots, I,$ then failures of ignorable treatment assignment given x_i can always be expressed in terms of a single unobserved covariate u_i with $0 \le u_i \le 1$ such that treatment assignment is ignorable given (x,u). This alternative to ignorable treatment assignment given x is the model for sensitivity analysis in Chapter 9, where $\Gamma < \infty$ implies $0 < \Pr(Z_i = 1|x_i, r_{Ti}, r_{Ci}) < 1$, for $i = 1, \ldots, I$.

34. The term "ignorable treatment assignment" is used informally for several mathematical conditions that express essentially the same idea. The term was first used in connection with Bayesian inference by Donald B. Rubin, "Bayesian inference for causal effects: The role of randomization," *Annals of Statistics* 6 (1978): 34–58. There was an era when some Bayesian statisticians said that randomization was not needed, but Rubin argues that a Bayesian who makes this claim is ignoring a factor in the likelihood that can only safely be ignored when randomization is used. In other words, the term "ignorable" is said in a tone of gentle reproach: you should only ignore how

treatments were assigned when the assignment is uninformative, as in a randomized experiment—you should only ignore treatment assignment when it is ignorable. Other terms in the literature that refer to the same or similar mathematical conditions are "no unmeasured confounders," "no unmeasured covariates," and "no hidden bias."

35. Rosenbaum and Rubin, "Central role of the propensity score," theorem 3.

6. Natural Experiments

1. Guido I. Imbens, Donald B. Rubin, and Bruce I. Sacerdote, "Estimating the effect of unearned income on labor earnings, savings and consumption: Evidence from a survey of lottery players," *American Economic Review* 91 (2001): 778–794.

2. Steven Stillman, John Gibson, David McKenzie, and Halahingano Rohorua, "Miserable migrants? Natural experiment evidence on international migration and objective and subjective well-being," *World Development* 65 (2015): 79–93.

3. Daniel E. Ho and Kosuke Imai, "Estimating causal effects of ballot order from a randomized natural experiment: The California alphabet lottery, 1978–2002," *Public Opinion Quarterly* 72 (2008): 216–240.

4. Daniel S. Nagin and G. Matthew Snodgrass, "The effect of incarceration on re-offending: Evidence from a natural experiment in Pennsylvania," *Journal of Quantitative Criminology* 29 (2013): 601–642.

5. Phillip Oreopoulus, "Long-run consequences of living in a poor neighborhood," *Quarterly Journal of Economics* 118 (2003): 1533–1575.

6. Alas, genes that are close together on the same chromosome tend to be transmitted together; they are said to be linked, or there is said to be linkage. Each gene and each part of a gene is randomized, but different genes and different parts of the same gene are not independently randomized, so some care and effort may be required, and some residual uncertainty may be present, when tracing a genetic effect to a particular aspect of a particular gene.

7. D. Curtis, "Use of siblings as controls in case-control association studies," *Annals of Human Genetics* 61 (1997): 319–333; Michael Boehnke and Carl D. Langefeld, "Genetic association mapping based on discordant sib pairs: The discordant-alleles test," *American Journal of Human Genetics* 62 (1998): 950–961; Richard S. Spielman and Warren J. Ewens, "A sibship test for linkage in the presence of association: The sib transmission / disequilibrium test," *American Journal of Human Genetics* 62 (1998): 450–458.

8. Darrin R. Lehman, Camille B. Wortman, and Allan F. Williams, "Long-term effects of losing a spouse or a child in a motor vehicle crash," *Journal of Personality and Social Psychology* 52 (1987): 218–231.

9. For several measures of depression, the difference in means was more than half the standard deviation, sometimes substantially more. This is a large effect, not easily

explained away by small biases due to nonrandom assignment, a topic that will be discussed in Chapter 9.

10. Lehman et al., "Long-term effects," 227–228.

11. Luke Keele and William Minozzi, "How much is Minnesota like Wisconsin? Assumptions and counterfactuals in causal inference with observational data," *Political Analysis* 21 (2013): 193–216.

12. As discussed in Chapter 7, this was done in a study by Garren J. Wintemute, Mona A. Wright, Christiana M. Drake, and James J. Beaumont, "Subsequent criminal activity among violent misdemeanants who seek to purchase handguns," *Journal of the American Medical Association* 285 (2001): 1019–1026. Jack McKillip spoke of "control outcomes"; see Jack McKillip, "Research without control groups," in *Methodological Issues in Applied Social Psychology,* ed. F. B. Bryant, 159–175 (New York: Plenum Press, 1992). For some general results about unaffected outcomes, see Paul R. Rosenbaum, *Observational Studies,* 2nd ed. (New York: Springer, 2002), chapter 6.

13. Erasmus, *The Collected Works of Erasmus,* vol. 34: *Adages* (Toronto: University of Toronto Press), 123.

14. Zena Stein, Mervyn Susser, Gerhart Saenger, and Francis Marolla, "Nutrition and mental performance: Prenatal exposure to the Dutch famine of 1944–1945," *Science* 178 (1972): 708–713.

15. For instance, see the following observational study and subsequent randomized experiment. Ewan Cameron and Linus Pauling, "Supplemental ascorbate in the supportive treatment of cancer: Prolongation of survival times in terminal human cancer," *Proceedings of the National Academy of Sciences* 73 (1976): 3685–3689; Charles G. Moertel, Thomas R. Fleming, Edward T. Creagan, Joseph Rubin, Michael J. O'Connell, and Matthew M. Ames, "High-dose vitamin C versus placebo in the treatment of patients with advanced cancer who have had no prior chemotherapy: a randomized double-blind comparison," *New England Journal of Medicine* 312 (1985): 137–141. These studies are discussed in Rosenbaum, *Observational Studies,* §1.2 and §4.4.3.

16. For instance, the following serious, scholarly studies of the effects of the minimum wage on employment reach very different conclusions. David Card and Alan B. Krueger, *Myth and Measurement: The New Economics of the Minimum Wage* (Princeton, NJ: Princeton University Press, 1995); David Neumark and William L. Wascher, *Minimum Wages* (Cambridge, MA: MIT Press, 2008). My personal favorite of these studies is David Card and Alan B. Krueger, "Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania," *American Economic Review* 84 (1994): 772–793.

17. Arthur L. Herbst, Howard Ulfelder, and David C. Poskanzer, "Adenocarcinoma of the vagina: Association of maternal stilbestrol therapy with tumor appearance in young women," *New England Journal of Medicine* 284 (1971): 878–881.

18. Rosenbaum, *Observational Studies,* 125–127.

19. To repeat from Chapter 5, note 33, treatment assignment is defined to be ignorable if $0 < \Pr(Z_i = 1 | x_i, r_{Ti}, r_{Ci}) = \Pr(Z_i = 1 | x_i) < 1$ for $i = 1, \ldots, I$. Here, $\pi_i = \Pr(Z_i = 1 | x_i, r_{Ti}, r_{Ci})$ and the propensity score is $\Pr(Z_i = 1 | x_i)$.

20. Even with lotteries, there are things to consider. One person buys two lottery tickets every week, year in and year out. Another person buys one lottery ticket once a year when the jackpot is huge and everyone is talking about it. Those people may differ in many ways—one hopes to win, the other wants to be part of the conversation—and the first person with many tickets is more likely to win.

21. Studies of disparities ask, sometimes pointedly, Why do two groups have different outcomes? Why do women earn less than men, on average? Studies of disparities use some of the tools used in observational studies, but they have a different structure. A study of a disparity is an attempt to accurately depict the mechanism by which the disparity is produced, but it does not contemplate the effects of, say, changing men into women. For an example consistent with my views about how disparities should be studied, see Jeffrey H. Silber, Paul R. Rosenbaum, Amy S. Clark, Bruce J. Giantonio, Richard N. Ross, Yun Teng, Min Wang, Bijan A. Niknam, Justin M. Ludwig, Wei Wang, Orit Even-Shoshan, and Kevin R. Fox, "Characteristics associated with differences in survival among black and white women with breast cancer," *Journal of the American Medical Association* 310 (2013): 389–397. For a related statistical method, see Paul R. Rosenbaum and Jeffrey H. Silber, "Using the exterior match to compare two entwined matched control groups," *American Statistician* 67 (2013): 67–75.

22. This point is emphasized by Richard Peto, M. Pike, Peter Armitage, Norman Breslow, David Cox, S. Howard, Nathan Mantel, K. McPherson, Julian Peto, and P. Smith, "Design and analysis of randomized clinical trials requiring prolonged observation of each patient," *British Journal of Cancer* 34 (1976): 585–612. They write, "A positive result is more likely, and a null result is more informative, if the main comparison is of only 2 treatments, these being as different as possible . . . It is the mark of good trial design that a null result, if it occurs, will be of interest" (590).

23. Mervyn Susser, "The challenge of causality: Human nutrition, brain development and mental performance," *Bulletin of the New York Academy of Medicine* 65 (1989): 1032–1049.

24. José R. Zubizarreta, Magdalena Cerdá, and Paul R. Rosenbaum, "Effect of the 2010 Chilean earthquake on posttraumatic stress: Reducing sensitivity to unmeasured bias through study design," *Epidemiology* 24 (2013): 79–87.

25. Y. Neria, A. Nandi, and S. Galea, "Post-traumatic stress disorder following disasters: A systematic review," *Psychological Medicine* 38 (2008): 467–480; Filip K. Arnberg, Ragnhildur Gudmundsdóttir, Agnieszka Butwicka, Fang Fang, Paul Lichtenstein, Christina M. Hultman, and Unnur A. Valdimarsdóttir, "Psychiatric disorders and suicide attempts in Swedish survivors of the 2004 southeast Asia tsunami: A 5 year matched cohort study," *Lancet Psychiatry* 2 (2015): 817–824.

26. Michael Rutter, "Resilience in the face of adversity: protective factors and resistance to psychiatric disorder," *British Journal of Psychiatry* 147 (1985): 598–611; G. A. Bonanno, "Loss, trauma, and human resilience: Have we underestimated the human capacity to thrive after extremely aversive events?," *American Psychologist* 59 (2004): 20–28.

27. Paul R. Rosenbaum, "The consequences of adjustment for a concomitant variable that has been affected by the treatment," *Journal of the Royal Statistical Society,* Series A, 147 (1984): 656–666.

28. Jonathan R. Davidson, S. W. Book, J. T. Colket, L. A. Tupler, S. Roth, D. David, M. Hertzberg, T. Mellman, J. C. Beckham, R. D. Smith, and R. M. Davison, "Assessment of a new self-rating scale for posttraumatic stress disorder," *Psychological Medicine* 27 (1997): 153–160.

29. José R. Zubizarreta, "Using mixed integer programming for matching in an observational study of kidney failure after surgery," *Journal of the American Statistical Association* 107 (2012): 1360–1371; José R. Zubizarreta, Ricardo D. Paredes, and Paul R. Rosenbaum, "Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in Chile," *Annals of Applied Statistics* 8 (2014): 204–231; Paul R. Rosenbaum and Donald B. Rubin, "Constructing a control group using multivariate matched sampling methods that incorporate the propensity score," *American Statistician* 39 (1985): 33–38; Paul R. Rosenbaum, "Optimal matching in observational studies," *Journal of the American Statistical Association* 84 (1989): 1024–1032; Paul R. Rosenbaum, Richard N. Ross, and Jeffrey H. Silber, "Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer," *Journal of the American Statistical Association* 102 (2007): 75–83.

30. Zubizarreta and colleagues did not cut the Davidson Trauma Scores into four categories, as in Tables 6.2 and 6.3; rather, I have done this to provide a compact display that is easy to examine and discuss. Their analysis distinguished all of the possible values, from 34 to 170, not four groups, but did so in a way that emphasized substantial symptoms. For discussion in the context of the Chilean earthquake, see Zubizarreta et al., "Effect of the 2010 Chilean earthquake"; for the general method, see Paul R. Rosenbaum, "Confidence intervals for uncommon but dramatic responses to treatment," *Biometrics* 63 (2007): 1164–1171.

7. Elaborate Theories

1. Charles Sanders Peirce, "Some consequences of four incapacities," *Journal of Speculative Philosophy* 2 (1868): 140–157, reprinted in *The Pragmatism Reader: From Peirce through the Present,* ed. R. B. Talisse and S. F. Aikin, 12–36 (Cambridge, MA: Harvard University Press, 2011), 13.

2. Franz Kafka, *The Blue Octavo Notebooks* (Cambridge: Exact Change, 1917).

3. William G. Cochran, "Planning of observational studies of human populations (with discussion)," *Journal of the Royal Statistical Society,* Series A, 128 (1965): 234–266, quotations from pp. 252–253.

4. David E. Morton, Alfred J. Saah, Stanley L. Silberg, Willis L. Owens, Mark A. Roberts, and Marylou D. Saah, "Lead absorption in children of employees in a lead-related industry," *American Journal of Epidemiology* 115 (1982): 549–555.

5. U.S. Centers for Disease Control and Prevention, "What do parents need to know to protect their children?," www.cdc.gov/nceh/lead/acclpp/blood_lead_levels .htm.

6. John W. Tukey, *Exploratory Data Analysis* (Reading, MA: Addison-Wesley, 1977), chapter 2.

7. Charles Sanders Peirce, "On selecting hypotheses," in *Collected Papers of Charles Sanders Peirce,* ed. C. Hartshorne and P. Weiss, 413–422 (Cambridge, MA: Harvard University Press, 1960), quotation from pp. 418–419.

8. Karl R. Popper, *The Logic of Scientific Discovery* (New York: Harper and Row, 1968), 112, 251, 267.

9. William G. Cochran, "Observational studies," in *Statistical Papers in Honor of George W. Snedecor,* ed. T. A. Bancroft, 70–90 (Ames: Iowa State University Press, 1972), reprinted with extensive commentary in Dylan S. Small, "Introduction to *Observational Studies* and the reprint of Cochran's paper 'Observational Studies' and Comments," *Observational Studies* 1 (2015): 124–240, http://obsstudies.org/files /cochran_and_comments.pdf. The importance of causal mechanisms in economic analysis is discussed and illustrated by Angus Deaton, "Understanding the mechanisms of economic development," *Journal of Economic Perspectives* 24 (2010) 3–16. For another view of causal mechanisms, see Tyler J. Vanderweele, *Explanation in Causal Inference* (New York: Oxford University Press, 2015).

10. Paul R. Rosenbaum, "Reasons for effects," *Chance* 18 (2005): 5–10; Paul R. Rosenbaum, "Some counterclaims undermine themselves in observational studies," *Journal of the American Statistical Association* 110 (2015): 1389–1398.

11. Richard Doll and Austin Bradford Hill, "The mortality of doctors in relation to their smoking habits," *British Medical Journal* 1 (1954): 1451–1455.

12. Ernest L. Wynder, Evarts A. Graham, and Adele B. Croninger, "Experimental production of carcinoma with cigarette tar," *Cancer Research* 13 (1953): 855–869.

13. Oscar Auerbach, A. P. Stout, Cuyler Hammond, and Lawrence Garfinkel, "Changes in bronchial epithelium in relation to cigarette smoking and in relation to lung cancer," *New England Journal of Medicine* 265 (1961): 253–267.

14. Nicholas Weller and Jeb Barnes, *Finding Pathways: Mixed-Method Research for Studying Causal Mechanisms* (New York: Cambridge University Press, 2014); Paul R. Rosenbaum and Jeffrey H. Silber, "Matching and thick description in an observational study of mortality after surgery," *Biostatistics* 2 (2001): 217–232.

15. Jerry A. Fodor, "The dogma that didn't bark," *Mind* 100 (1991): 201–220, quotation from pp. 202–203.

16. Garren J. Wintemute, Mona A. Wright, Christiana M. Drake, and James J. Beaumont, "Subsequent criminal activity among violent misdemeanants who seek to purchase handguns," *Journal of the American Medical Association* 285 (2001): 1019–1026.

17. Paul R. Rosenbaum, "From association to causation in observational studies: The role of tests of strongly ignorable treatment assignment," *Journal of the American Statistical Association* 79 (1984): 41–48; Paul R. Rosenbaum, "On permutation tests for hidden biases in observational studies," *Annals of Statistics* 17 (1989): 643–653; Paul R. Rosenbaum, *Observational Studies,* 2nd ed. (New York: Springer, 2002), chapters 6–8.

18. Susan Haack, *Evidence and Inquiry* (Oxford: Blackwell, 1995), 81–82.

19. Haack speaks of "independent security" for a claim that is secure when certain evidence or reasons are set aside. In statistics, the word "independent" is heavily committed to another use. An alternative term for this important concept is "demarcated security," meaning secure when certain specific demarcated evidence or reasons are set aside.

20. Homer, *The Iliad of Homer,* trans. Richard Lattimore (Chicago: University of Chicago Press, 2011), 139.

21. Carrie L. Randall, "Alcohol and pregnancy: Highlights from three decades of research," *Journal of Studies on Alcohol* 62 (2001): 554–561.

22. P. Lemoine, H. Harousseau, J. P. Borteyru, and J. C. Menuet, "Les enfants de parents alcooliques: Anomalies observées: A propos de 127 cas," *Ouest médical* 21 (1968): 476–482; Kenneth L. Jones, David W. Smith, Christy N. Ulleland, and Ann P. Streissguth, "Pattern of malformation in offspring of chronic alcoholic mothers," *Lancet* 2 (1973): 1267–1271. The subsequent quote is from Jones et al., "Pattern of malformation in offspring."

23. Randall, "Alcohol and pregnancy," 555.

24. Notice that the case series selected patients who had both high levels of exposure to alcohol and certain clinical abnormalities, in effect presuming rather than observing a connection between exposure and abnormality. For discussion of case series, see Olaf M. Dekkers, Matthias Egger, Douglas G. Altman, and Jan P. Vandenbroucke, "Distinguishing case-series from cohort studies," *Annals of Internal Medicine* 156 (2012): 37–40. They write, "A case-series may be a study that samples patients with both a specific outcome and a specific exposure . . . Whereas a cohort study, in principle, enables the calculation of an absolute risk or a rate for the outcome, such a calculation is not possible in a case-series."

25. Beatrice Larroque, Monique Kaminski, Philippe Dehaene, Damien Subtil, Marie-Jo Delfosse, and Denis Querleu, "Moderate prenatal alcohol exposure and psy-

chomotor development at preschool age," *American Journal of Public Health* 85 (1995): 1654–1661.

26. Sarah N. Mattson, Nicole Crocker, and Tanya T. Nguyen. "Fetal alcohol spectrum disorders: neuropsychological and behavioral features." *Neuropsychology Review* 21 (2011): 81–101. A study expressing some skepticism about the effects of moderate alcohol consumption found, among other things, that the father's use of alcohol also predicted the child's IQ, and that adjustments for covariates such as parental education reduced if not eliminated the association between IQ and moderate prenatal alcohol exposure; see Rosa Alati, John MacLeod, Matthew Hickman, Kapil Sayal, Margaret May, George Davey Smith, and Debbie A. Lawlor, "Intrauterine exposure to alcohol and tobacco use and childhood IQ: Findings from a parental-offspring comparison within the Avon Longitudinal Study of Parents and Children," *Pediatric Research* 64 (2008): 659–666.

27. Kathleen K. Sulik, Malcolm C. Johnston, and Mary A. Webb, "Fetal alcohol syndrome: Embryogenesis in a mouse model," *Science* 214 (1981): 936–938.

28. Chrysanthy Ikonomidou, Petra Bittigau, Masahiko J. Ishimaru, David F. Wozniak, Christian Koch, Kerstin Genz, Madelon T. Price, Vanya Stefovska, Friederike Horster, Tanya Tenkova, Krikor Dikranian, and John W. Olney, "Ethanol-induced apoptotic neurodegeneration and fetal alcohol syndrome," *Science* 287 (2000): 1056–1060.

29. Mary L. Schneider, Colleen F. Moore, and Gary W. Kraemer, "Moderate alcohol during pregnancy: Learning and behavior in adolescent rhesus monkeys," *Alcoholism: Clinical and Experimental Research* 25 (2001): 1383–1392.

30. Timothy A. Cudd, "Animal model systems for the study of alcohol teratology," *Experimental Biology and Medicine* 230 (2005): 389–393.

31. National Institute on Alcohol Abuse and Alcoholism, *Tenth Special Report to the US Congress on Alcohol and Health* (Bethesda, MD: Department of Health and Human Services, 2000), 287–289.

32. Debbie A. Lawlor, Jake M. Najman, G. David Batty, Michael J. O'Callaghan, Gail M. Williams, and William Bor, "Early life predictors of childhood intelligence: findings from the Mater-University study of pregnancy and its outcomes," *Paediatric and Perinatal Epidemiology* 20 (2006): 148–162.

33. National Institute on Alcohol Abuse and Alcoholism, *Tenth Special Report,* 283.

34. Ludwig Wittgenstein, *On Certainty* (New York: Harper, 1972), no. 17.

35. Ron Gray, Raja A. S. Mukherjee, and Michael Rutter, "Alcohol consumption during pregnancy and its effects on neurodevelopment: What is known and what remains unknown," *Addiction* 104 (2009): 1270–1273.

36. Centers for Disease Control and Prevention (CDC), "Alcohol and pregnancy," www.cdc.gov/vitalsigns/fasd/; and CDC, "More than 3 million US women at risk for

alcohol-exposed pregnancy," www.cdc.gov/media/releases/2016/p0202-alcohol-exposed
-pregnancy.html.

37. Thomas Nagel, *The Last Word* (New York: Oxford University Press), 26, 61,
and 91.

38. Nicholas Rescher, *Pluralism: Against the Demand for Consensus* (New York: Ox-
ford University Press, 1995).

39. Paul R. Rosenbaum, "Replicating effects and biases," *American Statistician* 55
(2001): 223–227.

40. Mervyn Susser, *Causal Thinking in the Health Sciences* (New York: Oxford,
1973), 148.

41. George Polya, "Heuristic reasoning and the theory of probability," *American
Mathematical Monthly* 48 (1941): 450–465.

42. A nontechnical discussion of evidence factors with references to the technical
literature is given in Paul R. Rosenbaum, "How to see more in observational studies:
Some new quasi-experimental devices," *Annual Review of Statistics and Its Applica-
tion* 2 (2015): 21–48.

43. Frank Wilcoxon, "Individual comparisons by ranking methods," *Biometrics* 1
(1945): 80–83; Myles Hollander Douglas A. Wolfe, and Eric Chicken, *Nonparametric
Statistical Methods* (New York: John Wiley, 2013), chapter 3. In R, one computes the
test from the 33 pair differences in lead levels:

```
> d
1     −3    14    23    −6    42    22     5    18    47
6     36    25    30    25    16    32
16    60    17    15     9     4     1    23     0    25
13     2     7    14    −9    −3
>wilcox.test(d)
V = 499, p-value = 1.155e-05
```

44. Maurice G. Kendall, "A new measure of rank correlation," *Biometrika* 30
(1938): 81–93. Hollander et al., *Nonparametric Statistical Methods,* chapter 8. In R,
one computes the test from the 33 pair differences in lead levels and the 33 exposure
levels (1 = low, 2 = medium, 3 = high):

```
e
3     3     3     3     3     3     3     3     3     3
3     3     3     3     3     3     3     3     3     2
2     2     2     2
2     1     1     1     1     1     1     1     1
>cor.test(d,e,method="kendall")
data:   d and e
z = 2.5613, p-value = 0.01043
```

```
alternative hypothesis: true tau is not equal to 0
sample estimates: tau 0.3613153
```

45. These properties are proved in Paul R. Rosenbaum, "Evidence factors in observational studies," *Biometrika* 97 (2010): 333–345, §5.1.

46. Fisher's method is a special case of the truncated product of *P*-values introduced by Dmitri V.Zaykin, Lev A. Zhivotovsky, Peter H. Westfall, and Bruce S. Weir, "Truncated product method for combining p-values," *Genetic Epidemiology* 22 (2002): 170–185. Fisher's method has truncation parameter 1, meaning no truncation, but truncation parameters below 1, perhaps 0.1, are better when the method is used in a sensitivity analysis; see Jesse Y. Hsu, Dylan S. Small, and Paul R. Rosenbaum, "Effect modification and design sensitivity in observational studies," *Journal of the American Statistical Association* 108 (2013): 135–148. In R, install the sensitivitymv package and type:

```
> library(sensitivitymv)
> truncatedP(c(0.01043,1.155e-05),trunc=1)
[1] 2.039726e-06
> truncatedP(c(0.01043,1.155e-05),trunc=.1)
[1] 1.701797e-06
```

47. Paul R. Rosenbaum, "Some approximate evidence factors in observational studies," *Journal of the American Statistical Association* 106 (2011): 285–295.

48. See, for instance, the examples in Rosenbaum, "How to see more in observational studies," "Evidence factors in observational studies," and "Some approximate evidence factors." For additional examples of evidence factors, see Kai Zhang, Dylan S. Small, Scott Lorch, Sindhu Srinivas, and Paul R. Rosenbaum, "Using split samples and evidence factors in an observational study of neonatal outcomes," *Journal of the American Statistical Association* 106 (2011): 511–524; José R. Zubizarreta, Mark Neuman, Jeffrey H. Silber, and Paul R. Rosenbaum, "Contrasting evidence within and between institutions that provide treatment in an observational study of alternate forms of anesthesia," *Journal of the American Statistical Association* 107 (2012): 901–915.

49. Ludwig Wittgenstein, *Philosophical Investigations* (New York: Macmillan, 1958), no. 265.

50. Peirce, "On selecting hypotheses," 419.

8. Quasi-experimental Devices

1. Imre Lakatos, *The Methodology of Scientific Research Programs* (New York: Cambridge University Press, 1978), 26. *"Ceteris paribus"* means "otherwise equal." To say that "a matched observational study estimates a treatment effect *ceteris paribus*"

is to say that this study would estimate the treatment effect if other unmeasured co-variates were equal in treated and control groups that have been matched for ob-served covariates.

2. Donald T. Campbell, "Factors relevant to the validity of experiments in social settings," *Psychological Bulletin* 54 (1957): 297–312. For an excellent modern textbook about quasi-experimental devices, see William R. Shadish, Thomas D. Cook, and Donald T. Campbell, *Experimental and Quasi-experimental Designs for Generalized Causal Inference* (Boston: Houghton Mifflin, 2002).

3. Donald T. Campbell and H. Laurence Ross, "The Connecticut crackdown on speeding: Time-series data in quasi-experimental analysis," *Law and Society Review* 3 (1968): 33–54.

4. Ernest Sosa, *A Virtue Epistemology: Apt Belief and Reflective Knowledge* (New York: Oxford University Press, 2007), 49.

5. Alfred R. Mele, *Self-Deception Unmasked* (Princeton, NJ: Princeton University Press, 2001), 26, 27, and 43.

6. Zena Stein, Mervyn Susser, Gerhart Saenger, and Francis Marolla, "Nutrition and mental performance: Prenatal exposure to the Dutch famine of 1944–1945," *Science* 178 (1972): 708–713.

7. Wayne A. Ray, Katherine T. Murray, Kathi Hall, Patrick G. Arbogast, and C. Michael Stein, "Azithromycin and the risk of cardiovascular death," *New England Journal of Medicine* 366 (2012): 1881–1890.

8. T. P. Van Staa, H. G. M. Leufkens, and C. Cooper, "Use of inhaled cortico-steroids and risk of fractures," *Journal of Bone and Mineral Research* 16 (2001): 581–588, quotation from p. 581. Van Staa and colleagues had previously claimed that oral corticosteroids, as distinct from inhaled corticosteroids, do cause fractures by causing osteoporosis; see T. P. Van Staa, H. G. M. Leufkens, L. Abenhaim, B. Zhang, and C. Cooper, "Use of oral corticosteroids and risk of fractures," *Journal of Bone and Mineral Research* 15 (2000): 993–1000.

9. Van Staa et al., "Use of inhaled corticosteroids," 586. They wrote: "Docu-mented prescription was a means of ensuring active registration at the practice. Therefore, we selected users of nonsystemic corticosteroids as controls." This is an instance of using a differential effect to remove a generic bias, as discussed in Chapter 12.

10. Frank B. Yoon, Haiden A. Huskamp, Alisa B. Busch, and Sharon-Lise T. Normand, "Using multiple control groups and matching to address unobserved biases in comparative effectiveness research: An observational study of the effectiveness of mental health parity," *Statistics in the Biosciences* 3 (2011): 63–78.

11. Donald T. Campbell, "Prospective: Artifact and control," in *Artifact in Behavioral Research,* ed. Robert Rosenthal and Ralph L. Rosnow, 264–286 (New York: Academic Press, 1969).

12. M. E. Bitterman, "Phyletic differences in learning," *American Psychologist* 20 (1965): 396–410.

13. This argument may be developed precisely to say that, under certain well-defined circumstances, systematic variation of u_i yields a consistent and unbiased test of an aspect of ignorable treatment assignment; see Paul R. Rosenbaum, "The role of a second control group in an observational study (with discussion)," *Statistical Science* 2 (1987): 292–306; and Paul R. Rosenbaum, *Observational Studies,* 2nd ed. (New York: Springer, 2002), chapter 8.

14. Campbell, "Prospective," 272.

15. Elizabeth A. Stuart and Donald B. Rubin, "Matching with multiple control groups with adjustment for group differences," *Journal of Educational and Behavioral Statistics* 33 (2008): 279–306; Samuel D. Pimentel, Dylan S. Small, and Paul R. Rosenbaum, "Constructed second control groups and attenuation of unmeasured biases," *Journal of the American Statistical Association* 111 (2016): 1157–1167.

16. Rosenbaum, "Role of a second control group"; Rosenbaum, *Observational Studies,* chapter 8.

17. See the discussion of equivalence tests in Chapter 3, or see Roger L. Berger and Jason C. Hsu, "Bioequivalence trials, intersection-union tests and equivalence confidence sets," *Statistical Science* 11 (1996): 283–319; and Jelle J. Goeman, Aldo Solari, and Theo Stijnen, "Three-sided hypothesis testing: Simultaneous testing of superiority, equivalence and inferiority," *Statistics in Medicine* 29 (2010): 2117–2125.

18. Yosef Hochberg and Ajit C. Tamhane, *Multiple Comparison Procedures* (New York: John Wiley, 1987).

19. A few of the many "quitting" strategies are described in the following works. These strategies are easy to use, but showing that they work is a technical topic. You should not try to invent a new "quitting" strategy unless you have the technical training required to do so. Ruth Marcus, Eric Peritz, and K. Ruben Gabriel, "On closed testing procedures with special reference to ordered analysis of variance," *Biometrika* 63 (1976): 655–660; Gary G. Koch and Stuart A. Gansky, "Statistical considerations for multiplicity in confirmatory protocols," *Drug Information Journal* 30 (1996): 523–534; Jason C. Hsu and Roger L. Berger, "Stepwise confidence intervals without multiplicity adjustment for dose—response and toxicity studies," *Journal of the American Statistical Association* 94 (1999): 468–482; Gerhard Hommel and Siegfried Kropf, "Tests for differentiation in gene expression using a data-driven order or weights for hypotheses," *Biometrical Journal* 47 (2005): 554–562, §3; Brian L. Wiens and Alexei Dmitrienko, "The fallback procedure for evaluating a single family of hypotheses," *Journal of Biopharmaceutical Statistics* 15 (2005): 929–942; Jelle J. Goeman and Aldo Solari, "The sequential rejection principle of familywise error control," *Annals of Statistics* (2010): 3782–3810.

20. This specific method is described in detail and illustrated in Paul R. Rosenbaum, "Testing hypotheses in order," *Biometrika* 95 (2008): 248–252. A different,

related approach was used by Yoon et al., "Using multiple control groups." A different approach to a related problem is developed by Michael Rosenblum, Han Liu, and En-Hsu Yen, "Optimal tests of treatment effects for the overall population and two subpopulations in randomized trials, using sparse linear programming," *Journal of the American Statistical Association* 109 (2014): 1216–1228.

21. George Polya, *How to Solve It* (Princeton, NJ: Princeton University Press, 2014), 117 and 232; George Polya, *Mathematics and Plausible Reasoning,* vol. 1: *Induction and Analogy in Mathematics* (Princeton, NJ: Princeton University Press, 1990), 28–29.

22. The literature uses the term "nonequivalent controls" rather than "counterpart." To my ear, "nonequivalent controls" is an oxymoron, a cow-horse. In an analogy, discernibly different things stand in the same relationship. In mathematics, a group of permutations, a group of matrices, and an abstract group defined by generators and relations may be isomorphic—that is, perfectly analogous—yet they are not the same; rather, they are perfect counterparts. With such perfect counterparts, you can perfectly predict what will happen in one system by examining what happens in its counterparts.

23. The situation described here is sometimes called a "nonequivalent control group design" and at other times called "the method of difference-in-differences," but both terms are also used to describe other kinds of comparisons in which both time and comparison groups play a role. Here, I am trying to emphasize one distinctive aspect of the situation: the counterparts are unambiguously not valid controls because they are definitely very different in ways that we cannot avoid and can easily see in the available data.

24. Bruce D. Meyer, W. Kip Viscusi, and David L. Durbin, "Worker's compensation and injury duration: Evidence from a natural experiment," *American Economic Review* 85 (1995): 322–340.

25. Meyer et al., "Worker's compensation, " figure 1.

26. These matched comparisons were not part of the original study. The comparison is limited to men, who comprise most of the data. Kentucky and Michigan were matched separately, as were higher earners and low earners, making four matched samples, reported here with Kentucky and Michigan combined. The matching used a propensity score, minimized a covariate distance, and employed an optimal subset, which affects the sample sizes. See Paul R. Rosenbaum, "Optimal matching of an optimally chosen subset in observational studies," *Journal of Computational and Graphical Statistics* 21 (2012): 57–71.

27. The logs are related by constant multipliers. For example, $\log_2(a) = \log_{10}(a) / \log_{10}(2)$.

28. The test and estimate are based on applying Wilcoxon's two sample test to the two unrelated groups of pair differences of log-durations in Figure 8.3. The associated estimate is the Hodges-Lehmann estimate. In R,

```
    Wilcoxon rank sum test with continuity correction
   W = 3514300, p-value = 6.178e-13
```

```
alternative hypothesis: true location shift is not
  equal to 0
95 percent confidence interval:
4.374038e-05  3.219662e-01
sample estimates:
difference in location
               0.2223606
> 2^0.2223606
[1] 1.166641
```

29. That is, some interactions can be removed by transformations. John W. Tukey, "One degree of freedom for non-additivity," *Biometrics* 5 (1949): 232–242.

30. Shadish et al., *Experimental and Quasi-experimental Designs.*

31. A distinct but related argument is developed in Paul R. Rosenbaum, "Stability in the absence of treatment," *Journal of the American Statistical Association* 96 (2001): 210–219.

32. Noel S. Weiss, "Can the 'specificity' of an association be rehabilitated as a basis for supporting a causal hypothesis?," *Epidemiology* 13 (2002): 6–8. For another example, see Kim D. Reynolds and Steven G. West, "A multiplist strategy for strengthening nonequivalent control group designs," *Evaluation Review* 11 (1987): 691–714.

33. Elizabeth Richardson Vigdor and James A. Mercy, "Do laws restricting access to firearms by domestic violence offenders prevent intimate partner homicide?," *Evaluation Review* 30 (2006): 313–346.

34. Vigdor and Mercy, "Do laws restricting access," 313.

35. Vigdor and Mercy, "Do laws restricting access," table 6.

36. General or theoretical discussions of the control outcomes examine the conditions under which they have a reasonable prospect of detecting unmeasured biases when biases are present. For general discussion of the method, see the following works: Paul R. Rosenbaum, "The role of known effects in observational studies," *Biometrics* 45 (1989): 557–569; Paul R. Rosenbaum, "On permutation tests for hidden biases in observational studies," *Annals of Statistics* 17 (1989): 643–653; Jack McKillip, "Research without control groups," in *Methodological Issues in Applied Social Psychology,* ed. F. B. Bryant, 159–175 (New York: Plenum Press, 1992); Rosenbaum, *Observational Studies,* chapter 6; Eric Tchetgen Tchetgen, "The control outcome calibration approach for causal inference with unobserved confounding," *American Journal of Epidemiology* 179 (2014): 633–640.

37. This strategy was first proposed by Donald L. Thistlethwaite and Donald T. Campbell, "Regression-discontinuity analysis: An alternative to the ex post facto experiment," *Journal of Educational Psychology* 51 (1960): 309. See also Jinyong Hahn, Petra Todd, and Wilbert Van der Klaauw, "Identification and estimation

of treatment effects with a regression-discontinuity design," *Econometrica* 69 (2001): 201–209; and Guido W. Imbens and Thomas Lemieux, "Regression discontinuity designs: A guide to practice," *Journal of Econometrics* 142 (2008): 615–635.

38. Joshua D. Angrist and Victor Lavy, "Using Maimonides' rule to estimate the effect of class size on student achievement," *Quarterly Journal of Economics* 114 (1999): 533–575.

39. Sandra E. Black, "Do better schools matter? Parental valuation of elementary education," *Quarterly Journal of Economics* 114 (1999): 577–599.

40. Luke Keele, Rocío Titiunik, and José R. Zubizarreta, "Enhancing a geographic regression discontinuity design through matching to estimate the effect of ballot initiatives on voter turnout," *Journal of the Royal Statistical Society,* Series A, 178 (2015): 223–239.

9. Sensitivity to Bias

1. Henry Petroski, *To Engineer Is Human: The Role of Failure in Successful Design* (New York: Vintage, 1992), 53.

2. John Dewey, "The pattern of inquiry: From *Logic: The Theory of Inquiry,*" in *The Essential Dewey,* vol. 2: *Ethics, Logic, Psychology,* ed. Larry A. Hickman and Thomas M. Alexander, 169–179 (Indianapolis: Indiana University Press, 1998), 171 (emphasis in original).

3. Donald B. Rubin, "The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials," *Statistics in Medicine* 26 (2007): 20–36.

4. Here, I am using the term "disconnected analyses" to distinguish this situation from formal multiple testing problems; see Alex Dmitrienko, Ajit C. Tamhane, and Frank Bretz, eds., *Multiple Testing Problems in Pharmaceutical Statistics* (Boca Raton, FL: Chapman and Hall / CRC, 2009); or Frank Bretz, Torsten Hothorn, and Peter Westfall, *Multiple Comparisons Using R* (Boca Raton, FL: Chapman and Hall / CRC, 2010).

5. Jerome Cornfield, William Haenszel, E. Cuyler Hammond, Abraham M. Lilienfeld, Michael B. Shimkin, and Ernst L, Wynder, "Smoking and lung cancer: recent evidence and a discussion of some questions," *Journal of the National Cancer Institute* 22 (1959): 173–203, reprinted with discussion in *International Journal of Epidemiology* 38 (2009): 1175–1191.

6. David R. Cox, "Commentary: Smoking and lung cancer: Reflections on a pioneering paper," *International Journal of Epidemiology* 38 (2009): 1192–1193; Jan P Vandenbroucke, "Commentary: 'Smoking and lung cancer'—the embryogenesis of modern epidemiology," *International Journal of Epidemiology* 38 (2009): 1193–1196;

Marcel Zwahlen, "Commentary: Cornfield on cigarette smoking and lung cancer and how to assess causality," *International Journal of Epidemiology* 38 (2009): 1197–1198; Joel B. Greenhouse, "Commentary: Cornfield, epidemiology and causality," *International Journal of Epidemiology* 38 (2009): 1199–1201. See also Samuel W. Greenhouse, "Jerome Cornfield's contributions to epidemiology," *Biometrics* 38, Suppl. (1982): 33–45; Binbing Yu and Joseph L. Gastwirth, "The use of the 'reverse Cornfield inequality' to assess the sensitivity of a non-significant association to an omitted variable," *Statistics in Medicine* 22 (2003): 3383–3401; Peng Ding and Tyler J. Vanderweele, "Generalized Cornfield conditions for the risk difference," *Biometrika* 101 (2014): 971–977.

7. Cornfield et al., "Smoking and lung cancer," 1186.

8. Greenhouse, "Commentary," 1200.

9. Ludwig Wittgenstein, *On Certainty* (New York: Harper and Row, 1969), no. 122.

10. Irwin D. J. Bross, "Statistical criticism," *Cancer* 13 (1960): 394–400.

11. Paul R. Rosenbaum and Donald B. Rubin, "Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome," *Journal of the Royal Statistical Society,* Series B, 45 (1983): 212–218; Takashi Yanagawa, "Case-control studies: Assessing the effect of a confounding factor," *Biometrika* 71 (1984): 191–194; Paul R. Rosenbaum, "Sensitivity analysis for certain permutation inferences in matched observational studies," *Biometrika* 74 (1987): 13–26; Charles F. Manski, "Nonparametric bounds on treatment effects," *American Economic Review* 80 (1990): 319–323; Joseph L. Gastwirth, "Methods for assessing the sensitivity of statistical comparisons used in Title VII cases to omitted variables," *Jurimetrics* 33 (1992): 19–34; Danyu Y. Lin, Bruce M. Psaty, and Richard A. Kronmal, "Assessing the sensitivity of regression results to unmeasured confounders in observational studies," *Biometrics* 54 (1998): 948–963; Charles F. Manski and Daniel S. Nagin, "Bounding disagreements about treatment effects: A case study of sentencing and recidivism," *Sociological Methodology* 28 (1998): 99–137; James M. Robins, Andrea Rotnitzky, and Daniel O. Scharfstein, "Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models," in *Statistical Models in Epidemiology, the Environment, and Clinical Trials,* ed. Elizabeth Halloran and D. Berry, 1–94 (New York: Springer, 2000); J. Copas and S. Eguchi, "Local sensitivity approximations for selectivity bias," *Journal of the Royal Statistical Society,* Series B, 63 (2001): 871–896; Guido W. Imbens, "Sensitivity to exogeneity assumptions in program evaluation," *American Economic Review* 93 (2003): 126–132; Thomas A. Diprete and Markus Gangl, "Assessing bias in the estimation of causal effects," *Sociological Methodology* 34 (2004): 271–310; Lawrence McCandless, P. Gustafson, and A. Levy, "Bayesian sensitivity analysis for unmeasured confounding in observational studies," *Statistics in Medicine* 26 (2007): 2331–2347; Carrie A. Hosman, Ben B. Hansen, and Paul W. Holland (2010), "The sensitivity of linear regression

coefficients' confidence limits to the omission of a confounder," *Annals of Applied Statistics* 4 (2010): 849–870; Jesse Y. Hsu and Dylan S. Small, "Calibrating sensitivity analyses to observed covariates in observational studies," *Biometrics* 69 (2013): 803–811.

For a few applications of sensitivity analysis in various fields, see the following works: Christopher S. Armstrong, Jennifer L. Blouin, and David F. Larcker, "The incentives for tax planning," *Journal of Accounting and Economics* 53 (2012): 391–411; E. Michael Foster, Elizabeth Wiley-Exley, and Leonard Bickman, "Old wine in new skins: the sensitivity of established findings to new methods," *Evaluation Review* 33 (2009): 281–306; Robert J. Glynn, Sebastian Schneeweiss, Philip S. Wang, Raia Levin, and Jerry Avorn, "Selective prescribing led to overestimation of the benefits of lipid-lowering drugs," *Journal of Clinical Epidemiology* 59 (2006): 819–828; Andrea Ichino, Fabrizia Mealli, and Tommaso Nannicini, "From temporary help jobs to permanent employment: What can we learn from matching estimators and their sensitivity?," *Journal of Applied Econometrics* 23 (2008): 305–327; Sharon-Lise T. Normand, Mary Beth Landrum, Edward Guadagnoli, John Z. Ayanian, Thomas J. Ryan, Paul D. Cleary, and Barbara J. McNeil, "Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: A matched analysis using propensity scores," *Journal of Clinical Epidemiology* 54 (2001): 387–398; Jeffrey H. Silber, Paul R. Rosenbaum, Richard N. Ross, Justin M. Ludwig, Wei Wang, Bijan A. Niknam, Nabanita Mukherjee, Phillip A. Saynisch, Orit Even-Shoshan, Rachel R. Kelz, and Lee A. Fleisher, "Template matching for auditing hospital cost and quality," *Health Services Research* 49 (2014): 1446–1474.

12. Paul R. Rosenbaum, *Observational Studies,* 2nd ed. (New York: Springer, 2002), chapter 4.

13. If i and j are two unrelated or independent people, then by definition

$$\theta_p = Pr(Z_i = 1 \text{ and } Z_j = 0 | Z_i + Z_j = 1) = \frac{\pi_i(1-\pi_j)}{\{\pi_i(1-\pi_j) + \pi_j(1-\pi_i)\}} ,$$

so that $\theta_p = \dfrac{1}{2}$ if $\pi_i = \pi_j$.

14. Use of Γ in various cases is extensively discussed in Rosenbaum, *Observational Studies,* chapter 4. Various R packages at `cran` perform sensitivity analyses, including my `sensitivitymv` and `sensitivitymw`, Luke Keele's `rbounds`, Dylan Small's `SensitivityCaseControl`, and `sensitivity2x2xk`. Markus Gangl's STATA module `rbounds` performs sensitivity analyses. See also Paul R. Rosenbaum, "Two R packages for sensitivity analysis in observational studies," *Observational Studies* 1 (2015): 1–17.

15. One might anticipate this from the formal definition of π_i in Chapter 5, note 33, as $\pi_i = Pr(Z_i = 1 | x_i, r_{Ti}, r_{Ci})$. That is, π_i has the outcome (r_{Ti}, r_{Ci}) built in from the

start, so in speaking about π_i we are already speaking about (r_{Ti}, r_{Ci}). In principle, with π_i so defined, we could *define u_i* to be π_i—this would be a mathematically coherent approach—but this definition of u_i would preclude some of the conversations that naturally occur in thinking about unmeasured covariates. We often wish to discuss a specific unobserved covariate, like alcohol consumption, as opposed to an unobserved covariate with a mathematical definition.

16. Gastwirth, "Methods for assessing the sensitivity."

17. The method described here in the last two columns of Table 9.1 is developed in Paul R. Rosenbaum and Jeffrey H. Silber, "Amplification of sensitivity analysis in matched observational studies," *Journal of the American Statistical Association* 104 (2009): 1398–1405. The article contains some technical detail that is important but not presented here in this book. The method alters, simplifies, and generalizes a related approach that is similar in spirit in Joseph L. Gastwirth, Abba M. Krieger, and Paul R. Rosenbaum, "Dual and simultaneous sensitivity analysis for matched pairs," *Biometrika* 85 (1998): 907–920. The method uses a semiparametric family of deformations of a symmetric distribution introduced in Douglas A. Wolfe, "A characterization of population weighted-symmetry and related results," *Journal of the American Statistical Association* 69 (1974): 819–822. The calculations are easy to do by hand, but the function `amplify` in the R package `sensitivitymv` automates the calculations. As an example, for the $\Gamma = 1.25$ row of Table 9.1:

```
> amplify(1.25,2)
2
2
```

18. The quantities Γ and Λ sound similar but are distinct: Γ refers to $\pi_i = \Pr(Z_i = 1 | x_i, r_{Ti}, r_{Ci})$ while Λ refers to $\Pr(Z_i = 1 | x_i, u_i)$.

19. Stated more precisely, two different statistical analyses based on two different models give identical answers when $\Gamma = (\Lambda\Delta + 1) / (\Lambda + \Delta)$. The derivation of this formula is given in Rosenbaum and Silber, "Amplification of sensitivity analysis."

20. David E. Morton, Alfred J. Saah, Stanley L. Silberg, Willis L. Owens, Mark A. Roberts, and Marylou D. Saah, "Lead absorption in children of employees in a lead-related industry," *American Journal of Epidemiology* 115 (1982): 549–555.

21. Frank Wilcoxon, "Individual comparisons by ranking methods," *Biometrics* 1 (1945): 80–83; Myles Hollander, Douglas A. Wolfe, and Eric Chicken, *Nonparametric Statistical Methods* (New York: John Wiley, 2013), chapter 3. Notes 43 and 44 of Chapter 7 include the data and computations in R.

22. For Wilcoxon's signed rank statistic, the elementary calculations are illustrated in Rosenbaum, *Observational Studies,* §4.3.3. For technical reasons, it is customary to perform two-sided randomization tests but to use one-sided *P*-values in sensitivity analyses, allowing the reader to double the one-sided *P*-value if the reader wishes to do this. I will follow this custom in the remainder of this chapter. The technical reasons are essentially the following. A randomization test may produce approximately

a Normal or Gaussian null distribution for a test statistic like Wilcoxon's signed rank statistic. Such a distribution puts equal probability in opposite tails. The sensitivity bounds are obtained by pushing much of the probability into one tail, say, the upper tail, thereby depleting the lower tail. For large Γ, doubling the one-sided P-value is quite conservative—there actually is not much probability left in the opposite tail—so that the size of the test is smaller than its nominal level. Of course, for $\Gamma = 1$, doubling the one-sided P-value is not conservative. The simplest approach is to perform a one-sided sensitivity analysis only if a two-sided randomization test rejects H_0, permitting an extremely cautious reader concerned with this issue to double the one-sided P-values from the sensitivity analysis.

23. That is, $(\Lambda, \Delta) = (8, 9.5)$ solves $4.4 = \Gamma = (\Lambda\Delta + 1)/(\Lambda + \Delta)$. In the `sensitivitymv` package in R:

```
> amplify(4.4,8)
8
9.5
```

24. Ludwig Wittgenstein, *On Certainty* (New York: Harper, 1972), no. 17.

25. For the lead data, with pair differences Y_p in `dif` from Chapter 7, note 43, the `sensitivitymw` package in R yields:

```
> senmw(dif,gamma=6,method="p")
$pval
[1] 0.0506652
```

For discussion of the `sensitivitymw` package, see Rosenbaum, "Two R packages." For discussion of the method used in this package, see Paul R. Rosenbaum, "Sensitivity analysis for m-estimates, tests, and confidence intervals in matched observational studies," *Biometrics* 63 (2007): 456–464; and Paul R. Rosenbaum, "Impact of multiple matched controls on design sensitivity in observational studies," *Biometrics* 69 (2013): 118–127. For discussion of the relative performance of different test statistics in sensitivity analyses, see Paul R. Rosenbaum, "Design sensitivity and efficiency in observational studies," *Journal of the American Statistical Association* 105 (2010): 692–702, and Paul R. Rosenbaum, "Bahadur efficiency of sensitivity analyses in observational studies," *Journal of the American Statistical Association* 110 (2015): 205–217.

26. See Rosenbaum, *Observational Studies,* §4.3.2, for an analysis of E. Cuyler Hammond, "Smoking in relation to mortality and morbidity: Findings in the first thirty-four months of follow-up in a prospective study started in 1959," *Journal of the National Cancer Institute* 32 (1964): 1161–1188.

27. The `sensitivitymw` package in R yields:

```
>senmwCI(dif,gamma=1,one.sided=FALSE)
$PointEstimate
minimum maximum
  15.41  15.41
```

```
$Confidence.Interval
minimum maximum
   10.14   21.60
>senmwCI(dif,gamma=2,one.sided=FALSE)
$PointEstimate
minimum maximum
   11.32   20.18
$Confidence.Interval
minimum maximum
    5.12   27.86
```

As with the *P*-value, results are less sensitive with a better test statistic; that is, the intervals are shorter and further from zero:

```
>senmwCI(dif,gamma=2,one.sided=FALSE,method="p")
$PointEstimate
minimum maximum
   11.36   19.13
$Confidence.Interval
minimum maximum
    6.2   28.2
```

28. Paul R. Rosenbaum, "The cross-cut statistic and its sensitivity to bias in observational studies with ordered doses of treatment," *Biometrics* 72 (2016) 175–183, §1.1.

29. Sharon K. Inouye, Sidney T. Bogardus Jr., Peter A. Charpentier, Linda Leo-Summers, Denise Acampora, Theodore R. Holford, and Leo M. Cooney Jr., "A multicomponent intervention to prevent delirium in hospitalized older patients," *New England Journal of Medicine* 340 (1999): 669–676.

30. In R, using `pbinom` from the `stats` package and `amplify` from the `sensitivitymw` package:

```
> pbinom(33,33+55,1/(1+1.14))
[1] 0.0510332
> amplify(1.14,1.5)
  1.5
  1.97
```

For discussion of sensitivity analyses for McNemar's test, see Rosenbaum, *Observational Studies,* §4.3.2. For discussion of sensitivity analyses for confidence intervals with binary outcomes, see Paul R. Rosenbaum, "Attributing effects to treatment in matched observational studies," *Journal of the American Statistical Association* 457 (2002): 183–192.

31. Vandenbroucke, "Commentary: 'Smoking and lung cancer.' "

32. See the example in Rosenbaum, *Observational Studies,* §1.2 and §4.4.3, for discussion of the following insensitive observational study and its subsequent correc-

tion by a randomized experiment: Ewan Cameron and Linus Pauling, "Supplemental ascorbate in the supportive treatment of cancer: Prolongation of survival times in terminal human cancer," *Proceedings of the National Academy of Sciences* 73 (1976): 3685–3689; and Charles G. Moertel, Thomas R. Fleming, Edward T. Creagan, Joseph Rubin, Michael J. O'Connell, and Matthew M. Ames, "High-dose vitamin C versus placebo in the treatment of patients with advanced cancer who have had no prior chemotherapy: a randomized double-blind comparison," *New England Journal of Medicine* 312 (1985): 137–141. The example is also discussed briefly in Paul R. Rosenbaum, "Sensitivity analysis for matching with multiple controls," *Biometrika* 75 (1988): 577–581.

33. Bruce D. Meyer, W. Kip Viscusi, and David L. Durbin, "Worker's compensation and injury duration: evidence from a natural experiment," *American Economic Review* 85 (1995): 322–340.

34. In the `sensitivitymw` package in R,

```
> senmw(high,gamma=1.3,method="p")
$pval
[1] 0.0437
> senmwCI(high,gamma=1.25,method="p")
$PointEstimate
minimum maximum
  0.47   1.95
$Confidence.Interval
minimum maximum
  0.12   Inf
```

35. José R. Zubizarreta, Magdalena Cerdá, and Paul R. Rosenbaum, "Effect of the 2010 Chilean earthquake on posttraumatic stress: Reducing sensitivity to unmeasured bias through study design," *Epidemiology* 24 (2013): 79–87.

36. Michael Rutter, "Resilience in the face of adversity: Protective factors and resistance to psychiatric disorder," *British Journal of Psychiatry* 147 (1985): 598–611; G. A. Bonanno, "Loss, trauma, and human resilience: Have we underestimated the human capacity to thrive after extremely aversive events?," *American Psychologist* 59 (2004): 20–28.

37. There are six mirror image pairs in Table 6.3, or six parallel comparisons, and there is one overall comparison with which we began, and finally there is the issue of one-sided versus two-sided tests. If I were pretending that I did not know where to look for a big effect (of course I knew where to look) and pretending that I made all those comparisons (which I did not), and I correct for seven tests and two-tails using the Bonferroni inequality, I still find that $8 \times 0.002334 = 0.018672 < 0.05$ at $\Gamma = 14$. So even if I pay an absurd price for data-snooping, it still pays to look at the extreme mirror image pairs in Table 6.3, because only by looking at the extreme pairs do I discover that that comparison is insensitive to very large biases. For some theory along

these lines, see Paul R. Rosenbaum, "Testing one hypothesis twice in observational studies," *Biometrika* 99 (2012): 763–774.

38. Zubizarreta et al., "Effect of the 2010 Chilean earthquake." The statistic S_m was proposed by Robert W. Stephenson, "A general class of one-sample nonparametric test statistics based on subsamples," *Journal of the American Statistical Association* 76 (1981): 960–966. A best test for the situation in which only some people are affected by treatment was developed by William J. Conover and David S. Salsburg, "Locally most powerful tests for detecting treatment effects when only a subset of patients can be expected to respond to treatment," *Biometrics* 44 (1988): 189–196. The close connection between these two tests, together with a sensitivity analysis is discussed in Paul R. Rosenbaum, "Confidence intervals for uncommon but dramatic responses to treatment," *Biometrics* 63 (2007): 1164–1171. Design sensitivities for S_m are calculated in Paul R. Rosenbaum, *Design of Observational Studies* (New York: Springer, 2010), chapter 16.

39. Can you use the data at hand to shop for a test statistic? Can you test Fisher's hypothesis several times using different test statistics? See Rosenbaum, "Testing one hypothesis twice"; and Paul R. Rosenbaum, "An exact adaptive test with superior design sensitivity in an observational study of treatments for ovarian cancer," *Annals of Applied Statistics* 6 (2012): 83–105.

10. Design Sensitivity

1. More precisely, if the Y_p were equal to a treatment effect of $\tau = 1$ plus errors from the long-tailed t-distribution with 2 degrees of freedom, then Wilcoxon's statistic has design sensitivity $\tilde{\Gamma} = 4.7$, while Stephenson's statistic, S_8, has lower design sensitivity $\tilde{\Gamma} = 3.7$. Other statistics beat Wilcoxon's statistic both for the Normal distribution and for the t-distribution with 2 degrees of freedom. The calculations in this section and similar calculations for other statistics are presented in table 3 of Paul R. Rosenbaum, "A new U-statistic with superior design sensitivity in observational studies," *Biometrics* 67 (2011): 1017–1027. The R packages `sensitivitymv` and `sensitivitymw` do not use these rank tests, but rather M-statistics, including the mean. Calculations of design sensitivities $\tilde{\Gamma}$ for paired M-tests including the mean are given in Corollary 1 of Paul R. Rosenbaum, "Impact of multiple matched controls on design sensitivity in observational studies," *Biometrics* 69 (2013): 118–127.

2. John Stuart Mill, *A System of Logic* (1843; reprinted in Mill, *The Collected Works of John Stuart Mill,* vol. 7: *A System of Logic Ratiocinative and Inductive, Being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation* [books 1–3], ed. John M. Robson, intro. R. F. McRae [Toronto: University of Toronto Press; London: Routledge and Kegan Paul, 1974]).

3. Ronald A. Fisher, *Design of Experiments* (Edinburgh: Oliver and Boyd, 1935, 1949), 18.

4. O. Ashenfelter and C. Rouse, "Income, schooling and ability: Evidence from a new sample of identical twins," *Quarterly Journal of Economics* 113 (1998): 253–284; Avshalom Caspi, Julia Kim-Cohen, Terrie E. Moffitt, Julia Morgan, Michael Rutter, Monica Polo-Tomas, and Alan Taylor, "Maternal expressed emotion predicts children's antisocial behavior problems: Using monozygotic-twin differences to identify environmental effects on behavioral development," *Developmental Psychology* 40 (2004): 149–161; A. Haapanen, M. Koskenvuo, J. Kaprio, Y. A. Kesäniemi, and K. Heikkilä, "Carotid arteriosclerosis in identical twins discordant for cigarette smoking," *Circulation* 80 (1989): 10–16.

5. David Card and Alan B. Krueger, "Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania," *American Economic Review* 84 (1994): 772–793. Burger King replicates its restaurants, a basic business strategy used by everyone from Carrefour to H&R Block; see Sidney G. Winter and Gabriel Szulanski, "Replication as strategy," *Organization Science* 12 (2001): 730–743. Business replicates are the genetically engineered mice of microeconomics.

6. This is demonstrated by computing the power of a sensitivity analysis. This pattern is not confined to the Normal distribution; it also is seen with long-tailed distributions like the logistic and the Cauchy distribution. See table 1 in Paul R. Rosenbaum, "Heterogeneity and causality: Unit heterogeneity and design sensitivity in observational studies," *American Statistician* 59 (2005): 147–152; or Paul R. Rosenbaum, *Design of Observational Studies* (New York: Springer, 2010), chapter 15.

7. The mean, $\overline{Y}$, has larger design sensitivity than Wilcoxon's statistic for Gaussian or Normal errors. There are statistics in the R package `sensitivitymw` with larger design sensitivity than $\overline{Y}$ for both Normal and long-tailed errors, such as the *t*-distribution with 3 degrees of freedom; see Rosenbaum, "Impact of multiple matched controls," table 3.

8. P. H. Wright and Leon S. Robertson, "Priorities for roadside hazard modification," *Traffic Engineering* 46 (1976): 24–30.

9. Malcolm Maclure, "The case-crossover design: A method for studying transient effects on the risk of acute events," *American Journal of Epidemiology* 133 (1991): 144–152; Sander Greenland, "A unified approach to the analysis of case-distribution (case-only) studies," *Statistics in Medicine* 18 (1999): 1–15.

10. José R. Zubizarreta, Ricardo D. Paredes, and Paul R. Rosenbaum, "Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in Chile," *Annals of Applied Statistics* 8 (2014): 204–231.

11. Zubizarreta et al., "Matching for balance," table 1.

12. Darrin R. Lehman, Camille B. Wortman, and Allan F. Williams, "Long-term effects of losing a spouse or a child in a motor vehicle crash," *Journal of Personality and Social Psychology* 52 (1987): 218–231.

13. L. Molineaux and G. Gramiccia, *The GARKI Project: Research on the Epidemiology and Control of Malaria in the Sudan Savanna of West Africa* (Geneva: World Health Organization, 1980); Jesse Y. Hsu, Dylan S. Small, and Paul R. Rosenbaum, "Effect modification and design sensitivity in observational studies," *Journal of the American Statistical Association* 108 (2013): 135–148.

14. José R. Zubizarreta, Magdalena Cerdá, and Paul R. Rosenbaum, "Effect of the 2010 Chilean earthquake on posttraumatic stress: Reducing sensitivity to unmeasured bias through study design," *Epidemiology* 24 (2013): 79–87.

15. Concerning the focus on intense versions of treatment, see Rosenbaum, *Design of Observational Studies,* §17.3; and Paul R. Rosenbaum, "Design sensitivity in observational studies," *Biometrika* 91 (2004): 153–164, table 3. Concerning effect modification, see Hsu et al., "Effect modification"; and Jesse Y. Hsu, José R. Zubizarreta, Dylan S. Small, and Paul R. Rosenbaum, "Strong control of the familywise error rate in observational studies that discover effect modification by exploratory methods," *Biometrika* 102 (2015): 767–782. Concerning treatments that affect some people but not everyone, see Zubizarreta et al., "Effect of the 2010 Chilean earthquake"; Rosenbaum, *Design of Observational Studies,* chapter 16; and references in Chapter 9, note 38. See also Peng Ding, Avi Feller, and Luke Miratrix, "Randomization inference for treatment effect variation," *Journal of the Royal Statistical Society,* Series B, 78 (2016): 655–671. For interesting results about design sensitivity with several outcomes, see Colin B. Fogarty and Dylan S. Small, "Sensitivity analysis for multiple comparisons in matched observational studies through quadratically constrained linear programming," *Journal of the American Statistical Association* 111 (2016): 1820–1830.

16. Ruth Heller, Paul R. Rosenbaum, and Dylan S. Small, "Split samples and design sensitivity in observational studies," *Journal of the American Statistical Association* 104 (2009): 1090–1101. For a discussion of statistical power with split samples, see David R. Cox, "A note on data-splitting for the evaluation of significance levels," *Biometrika* 62 (1975): 441–444. A more complex approach that uses all of the data is discussed in Paul R. Rosenbaum, "Testing one hypothesis twice in observational studies," *Biometrika* 99 (2012): 763–774.

17. Austin Bradford Hill, "The environment and disease: Association or causation?," *Proceedings of the Royal Society of Medicine* 58 (1965): 295–300.

18. Unlike criteria, it is commonplace to weigh one consideration against another, to set a consideration aside as not relevant in a particular context, and to weigh an old consideration against a new consideration never considered before. Hill was explicit about weighing considerations. For instance, in discussing one consideration,

plausibility, he wrote, "Plausibility: It will be helpful if the causation we suspect is biologically plausible. But this is a feature I am convinced we cannot demand. What is biologically plausible depends upon the biological knowledge of the day" ("Environment and disease," 298). For a discussion of "plausibility," see Douglas L. Weed and Stephen D. Hursting, "Biologic plausibility in causal inference: Current method and practice," *American Journal of Epidemiology* 147 (1998): 415–425.

Others have expressed skepticism about the widespread view that Hill intended to offer "criteria for causality," rather than considerations possibly relevant in reaching a thoughtful judgement. For instance, Kenneth Rothman and Sander Greenland write that Hill "disagreed that any 'hard-and-fast rules of evidence' existed by which to judge causation," and they then show though examples how mechanical application of "causal criteria" would have produced mistaken conclusions in particular instances; see Kenneth J. Rothman and Sander Greenland, "Causation and causal inference in epidemiology," *American Journal of Public Health* 95, Suppl. 1 (2005): S144–S150.

19. Hill, "Environment and disease," 298.

20. Nathan Mantel, "Chi-square tests with one degree of freedom; extensions of the Mantel-Haenszel procedure," *Journal of the American Statistical Association* 58 (1963): 690–700; M. W. Birch, "The detection of partial association, II: The general case," *Journal of the Royal Statistical Society,* Series B, 27 (1965): 111–124; Paul R. Rosenbaum, "Does a dose–response relationship reduce sensitivity to hidden bias?," *Biostatistics* 4 (2003): 1–10.

21. Rosenbaum, *Design of Observational Studies,* §17.3; Rosenbaum, "Design sensitivity," table 3.

22. Allan Donner and Neil Klar, *Design and Analysis of Cluster Randomization Trials in Health Research* (New York: John Wiley and Sons, 2010).

23. See Martha L. Bruce, Thomas R. Ten Have, Charles F. Reynolds III, Ira I. Katz, Herbert C. Schulberg, Benoit H. Mulsant, Gregory K. Brown, Gail J. McAvay, Jane L. Pearson, and George S. Alexopoulos, "Reducing suicidal ideation and depressive symptoms in depressed older primary care patients: A randomized controlled trial," *Journal of the American Medical Association* 291 (2004): 1081–1091; Dylan S. Small, Thomas R. Ten Have, and Paul R. Rosenbaum, "Randomization inference in a group–randomized trial of treatments for depression: Covariate adjustment, noncompliance, and quantile effects," *Journal of the American Statistical Association* 103 (2008): 271–279.

24. Carlo Del Ninno, Paul A. Dorosh, Lisa C. Smith, and Dilip K. Roy, "The 1998 floods in Bangladesh: Disaster impacts, household coping strategies and response," Research Report 122 (Washington, DC: International Food Policy Research Institute, 2001). This study supplied the example in Ben B. Hansen, Paul R. Rosenbaum, and Dylan S. Small, "Clustered treatment assignments and sensitivity to unmeasured biases in observational studies," *Journal of the American Statistical Association,* 109 (2014): 133–144. See also Ben B. Hansen and Jake Bowers, "Attributing effects to a

cluster-randomized get-out-the-vote campaign," *Journal of the American Statistical Association* 104 (2009): 873–885.

25. That is, suppose the intracluster correlation of outcomes under control is < 1.

26. Hansen et al., "Clustered treatment assignments," tables 2, 3, and 4.

27. Luke Keele and Jose R. Zubizarreta, "Optimal multilevel matching in clustered observational studies: A case study of the school voucher system in Chile," *Journal of the American Statistical Association* (2017, forthcoming); Samuel D. Pimentel, Lindsay Page, Matthew Lenard, and Luke Keele, "Optimal multilevel matching using network flows: An application to a summer reading intervention," preprint (2015). These algorithms are examples of dynamic programming.

28. Many issues affect design sensitivity. Some additional issues affecting design sensitivity are discussed in Chapter 11. A broad discussion of design sensitivity is the topic of part III of Rosenbaum, *Design of Observational Studies.*

11. Matching Techniques

1. Recall from Chapter 5, note 33, that this approach applies generally in the following technical sense. The propensity score, λ_x or $\lambda(x)$, is the conditional probability of treatment, $Z=1$, given the observed covariates, $\lambda(x) = \Pr(Z=1|x)$. That is, $\lambda(x)$ is the average of π_i over all individuals i who happen to have observed covariate x_i equal to a specific value x. Here, π is $\Pr(Z=1|x,r_T,r_C)$, so that, written concisely but technically, $\lambda(x) = \Pr(Z=1|x) = E\{\Pr(Z=1|x,r_T,r_C)|x\}$. By definition, treatment assignment is ignorable given x if $0 < \lambda(x_i) = \pi_i < 1$ for all i, or equivalently if $\Pr(Z=1|x) = \Pr(Z=1|x,r_T,r_C)$ and $0 < \Pr(Z=1|x) < 1$. Whenever the treatment assignment is not ignorable given x there always exists a scalar unobserved covariate u with $0 \le u \le 1$ such that $\Pr(Z=1|x,u) = \Pr(Z=1|x,u,r_T,r_C)$, namely, $u = \Pr(Z=1|x,r_T,r_C)$. This structure expresses departures from ignorable treatment assignment given x in terms of a single omitted covariate u with $0 \le u \le 1$, and it is compatible with the sensitivity analysis in Chapter 9. In Chapter 9, any Γ with $1 \le \Gamma < \infty$ entails ignorable treatment assignment given (x,u); that is, $0 < \Pr(Z=1|x,u) = \Pr(Z=1|x,u,r_T,r_C) < 1$. In that sense, the sensitivity analysis in Chapter 9 is the sensitivity analysis derived directly from propensity scores and ignorable assignment, with no added structures.

2. Jeffrey H. Silber, Paul R. Rosenbaum, Matthew D. McHugh, Justin M. Ludwig, Herbert L. Smith, Bijan A. Niknam, Orit Even-Shoshan, Lee A. Fleisher, Rachel R. Kelz, and Linda H. Aiken, "Comparison of the value of nursing work environments in hospitals across different levels of patient risk," *JAMA Surgery* 151 (2016): 527–536.

3. Linda H. Aiken, Donna S. Havens, and Douglas M. Sloane, "The magnet nursing services recognition program," *American Journal of Nursing* 100 (2000): 26–35.

4. This causal question is about what happens to patients when they make a different choice among existing hospitals, and it is a practical question facing patients and insurers. The counterfactual is, What would happen to Harry if he went to a different hospital? It is not remotely the same question as asking whether changing the nursing environment at a hospital would change the outcomes produced by that hospital. It is not remotely the same question as asking whether hiring more nursing improves the performance of a given hospital—that is a different counterfactual, one not addressed by the study of Silber and colleagues. Silber and colleagues observed many patients choosing among existing hospitals; they did not observe many hospitals actively improving their nursing environments. The 35 focal hospitals differed from the 293 control hospitals in many ways: the defining characteristic of superior nursing environments picked out hospitals that also had more medical residents per bed and advanced capabilities such as a burn center or the ability to perform bone marrow transplantation. It would not be surprising if the best surgeons and anesthesiologists prefer to work at hospitals with superior nursing environments, so differences in staffing between focal and control hospitals are unlikely to be confined to the nurses. The focal and control hospitals are two types of existing hospitals, and we are looking at the effect on Harry of going to a hospital of one type or the other. That is one interesting question, but it should not be confused with other interesting questions.

5. The odds ratio from McNemar's test was 0.79 with a 95% confidence interval of 0.73 to 0.86. To explain this difference in mortality as a bias from failing to match for an unobserved covariate u_i, the covariate would need to double the risk of death and be 50% more common among control hospitals (Silber et al., "Comparison of the value of nursing," table 3 and appendix 10).

6. Silber et al., "Comparison of the value of nursing," tables 3 and 5.

7. Silber et al., "Comparison of the value of nursing," table 3. Here is an example of the difficulty in evaluating payments. Medicare pays a slightly higher amount for the same surgical procedure if the procedure is performed in a place where costs are higher, say New York City. Focal hospitals are overrepresented in some large cities, say New York City. It is unclear whether this geographic adjustment to payments should be viewed as part of what Medicare pays for a focal hospital, or part of what Medicare pays to provide health care in expensive locations. After all, there are many control hospitals in New York City, and these control hospitals also receive the geographic adjustment to payments. There are several similar issues of this kind.

8. Silber et al., "Comparison of the value of nursing," appendix 1.

9. Letter to Robert Louis Stevenson, January 12, 1891, in Henry James, *The Letters of Henry James,* vol. 1 (New York: Charles Scribner's Sons, 1920), 174–179.

10. For instance, in Silber et al., "Comparison of the value of nursing," table 2 in the body of the article described the balance for 14 covariates, and appendix 8 described, in greater detail, the balance for 172 covariates. For example, the appendix

showed that the baseline estimated risk of death was similar at focal and control hospitals, so matching did little to change this. Also, the appendix showed how matching had altered the distribution of the 130 surgical procedures.

11. In matching with a variable number of controls or in full matching—described later in this chapter—matching may reduce the weight attached to some controls rather than excluding these controls. In this case, the table may show how the change in weights has changed the populations. For example, if in the unmatched treated and control populations, treated individuals are younger than controls, then matching with variable numbers of controls may reduce the weight attached to older controls so that the weighted control population is younger than the unweighted control population. Generally, matching with a variable number of controls may both exclude some controls and change the weights attached to those who remain, and full matching may do this in the treated group as well. A table showing how these changes have changed the distribution of observed covariates x_i is often helpful in understanding what biases in x_i existed before matching and how matching has removed them.

12. See, for instance, table 1 in Peter P. Reese, R. D. Bloom, Harvey I. Feldman, Paul R. Rosenbaum, Wei Wang, Phillip Saynisch, N. M. Tarsi, Nabanita Mukherjee, A. X. Garg, A. Mussell, Justine Shults, Orit Even-Shoshan, R. R. Townsend, and Jeffrey H. Silber, "Mortality and cardiovascular disease among older live kidney donors," *American Journal of Transplantation* 20 (2014): 1–9. See also Silber et al., "Comparison of the value of nursing," appendix 8.

13. As discussed in Chapter 10, in the section "Can One Search for an Insensitive Finding?" and its endnotes, a single, simple, prespecified planned analysis can be made compatible with some exploratory work. This is sometimes possible using specialized techniques. However, the most direct approach is to (i) split the sample at random into a 10% planning sample and a 90% analysis sample, (ii) build, in any way at all, a simple plan for the study using the 10% planning sample, (iii) discard the 10% planning sample, and (iv) implement the simple plan derived from the planning sample using the 90% analysis sample. See Ruth Heller, Paul R. Rosenbaum, and Dylan S. Small, "Split samples and design sensitivity in observational studies," *Journal of the American Statistical Association* 104 (2009): 1090–1101; and David R. Cox, "A note on data-splitting for the evaluation of significance levels," *Biometrika* 62 (1975): 441–444.

14. See Chapter 8, specifically the section "Organizing Analysis with Two Control Groups" and its endnotes.

15. Jesse Y. Hsu, José R. Zubizarreta, Dylan S. Small, and Paul R. Rosenbaum, "Strong control of the familywise error rate in observational studies that discover effect modification by exploratory methods," *Biometrika* 102 (2015): 767–782.

16. Samuel D. Pimentel, Dylan S. Small, and Paul R. Rosenbaum, "Constructed second control groups and attenuation of unmeasured biases," *Journal of the American Statistical Association* 111 (2016): 1157–1167.

17. Peter M. Aronow and Cyrus Samii, "Does regression produce representative estimates of causal effects?," *American Journal of Political Science* 60 (2016): 250–267. A related point is made in a different context by Andrew Gelman and Guido Imbens, "Why high-order polynomials should not be used in regression discontinuity designs," *National Bureau of Economic Research Working Papers,* 2014, #w20405. Both papers argue that certain model-based adjustments can allow some treated or control individuals to have too much weight in the analysis and that this is not obvious from the analysis itself. For instance, a control whose x is substantially unlike the x's for all treated individuals is not much of a control, but may exert high leverage in a model-based adjustment.

18. John Stuart Mill, *On Liberty* and *The Subjection of Women* (New York: Penguin, 2007), 38.

19. We should be cautious about assigning much weight to counterarguments that are exceedingly vague. It is not easy to explain just what is wrong with this sort of vagueness, just what is wrong with the vague statement: "That is not enough evidence." J. L. Austin expressed the matter well: "If you say 'That's not enough,' then you must have in mind some more or less definite lack . . . Enough is enough: it doesn't mean everything . . . The wile of the metaphysician consists in asking 'Is it a real table?' (a kind of object which has no obvious way of being phony) and not specifying or limiting what may be wrong with it, so that I feel at a loss 'how to prove' it is a real one." John L. Austin, *Philosophical Papers* (New York: Oxford University Press, 1979), 84–87.

20. A better strategy uses a penalty function. The distance between i and j is not penalized if $|\lambda(x_i) - \lambda(x_j)| \leq \kappa$, but if $|\lambda(x_i) - \lambda(x_j)| > \kappa$ then the distance between i and j is increased by adding a large quantity, the quantity growing larger as $|\lambda(x_i) - \lambda(x_j)|$ increases. The advantage of a penalty function is that if it is not possible to match so that $|\lambda(x_i) - \lambda(x_j)| \leq \kappa$ for all matched pairs, the match will let just a few pairs have $|\lambda(x_i) - \lambda(x_j)| > \kappa$ by just a small amount. For discussion of penalty functions in matching, see Paul R. Rosenbaum, *Design of Observational Studies* (New York: Springer, 2010), chapter 8.

21. Prasanta Chandra Mahalanobis, "On the generalized distance in statistics," *Proceedings of the National Institute of Sciences* (Calcutta) 2 (1936): 49–55. For discussion of the small adjustments needed to obtain a robust Mahalanobis distance, see Rosenbaum, *Design of Observational Studies,* chapter 8.

22. This distance is (essentially) the one recommended in Paul R. Rosenbaum and Donald B. Rubin, "Constructing a control group using multivariate matched sampling methods that incorporate the propensity score," *American Statistician* 39 (1985): 33–38.

23. Ben B. Hansen, "The prognostic analogue of the propensity score," *Biometrika* 95 (2008): 481–488. One cannot safely estimate the weights for a prognostic score from the current data, so the weights should come from a source not used as the

basis for the study's conclusions. The simplest approach uses published weights for some standard score, such as the APACHE II score or the Charlson Comorbidity Index.

Mike Baiocchi proposed a clever and practical device for safely building a prognostic score from data discarded by matching. One first matches without using a prognostic score. Some controls are not selected by matching. The prognostic score is estimated from these discarded controls. The pairs in the original match are broken, and these same individuals are paired a second time, now including the prognostic score estimated from the unused controls. Covariate balance is unaltered because it ignores who is matched to whom, but the new pairs may better predict the outcome, reducing heterogeneity in Y_i, thereby possibly increasing insensitivity to unmeasured covariates. See Mike Baiocchi, "Methodologies for observational studies of health care policy" (PhD diss., Department of Statistics, Wharton School of the University of Pennsylvania, 2011). When several outcomes are of interest, a single match can control for several externally estimated prognostic scores, as illustrated with five prognostic scores by Jeffrey H. Silber, Paul R. Rosenbaum, Richard N. Ross, Justin M. Ludwig, Wei Wang, Bijan A. Niknam, Alexander S. Hill, Orit Even-Shoshan, Rachel R. Kelz, and Lee A. Fleisher, "Indirect standardization matching: Assessing specific advantage and risk synergy," *Health Services Research* 51 (2016): 2330–2357.

24. For a matching device—strength k matching—that permits a more extensive search for effect modifiers, see Hsu et al., "Strong control of the familywise error rate."

25. Harold W. Kuhn, "The Hungarian method for the assignment problem," *Naval Research Logistics Quarterly* 2 (1955): 83–97; Dimitri P. Bertsekas, "A new algorithm for the assignment problem," *Mathematical Programming* 21 (1981): 152–171; Rainer E. Burkard, Mauro Dell'Amico, and Silvano Martello, *Assignment Problems* (Philadelphia: Society for Industrial and Applied Mathematics, 2009). Use of optimal matching algorithms in observational studies is discussed in Paul R. Rosenbaum, "Optimal matching for observational studies," *Journal of the American Statistical Association* 84 (1989): 1024–1032.

26. Paul R. Rosenbaum, Richard N. Ross, and Jeffrey H. Silber, "Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer," *Journal of the American Statistical Association* 102 (2007): 75–83; José R. Zubizarreta, "Using mixed integer programming for matching in an observational study of kidney failure after surgery," *Journal of the American Statistical Association* 107 (2012): 1360–1371.

27. This is illustrated by comparing a match with fine balance and another match without fine balance in tables 5 and 6 of José R. Zubizarreta, Caroline E. Reinke, Rachel R. Kelz, Jeffrey H. Silber, and Paul R. Rosenbaum, "Matching for several sparse nominal variables in a case-control study of readmission following surgery,"

American Statistician 65 (2011): 229–238. For a few applications of matching with fine balance, see the following publications: Jeffrey H. Silber, Paul R. Rosenbaum, Daniel Polsky, Richard N. Ross, Orit Even-Shoshan, J. Sanford Schwartz, Katrina A. Armstrong, and Thomas C. Randall, "Does ovarian cancer treatment and survival differ by the specialty providing chemotherapy?," *Journal of Clinical Oncology* 25 (2007): 1169–1175; Mark D. Neuman, Paul R. Rosenbaum, Justin M. Ludwig, Jose R. Zubizarreta, and Jeffrey H. Silber, "Anesthesia technique, mortality, and length of stay after hip fracture surgery," *Journal of the American Medical Association* 311 (2014): 2508–2517; Rachel R. Kelz, Caroline E. Reinke, José R. Zubizarreta, Min Wang, Philip Saynisch, Orit Even-Shoshan, Peter P. Reese, Lee A. Fleisher, and Jeffrey H. Silber, "Acute kidney injury, renal function, and the elderly obese surgical patient: a matched case-control study," *Annals of Surgery* 258 (2013): 359–363.

28. Dan Yang, Dylan S. Small, Jeffrey H. Silber, and Paul R. Rosenbaum, "Optimal matching with minimal deviation from fine balance in a study of obesity and surgical outcomes," *Biometrics* 68 (2012): 628–636.

29. Samuel D. Pimentel, Rachel R. Kelz, Jeffrey H. Silber, and Paul R. Rosenbaum, "Large, sparse optimal matching with refined covariate balance in an observational study of the health outcomes produced by new surgeons," *Journal of the American Statistical Association* 110 (2015): 515–527.

30. Herbert L. Smith, "Matching with multiple controls to estimate treatment effects in observational studies," *Sociological Methodology* 27 (1997): 325–353.

31. The model assumes ignorable treatment assignment, an additive matched set parameter that is removed by taking differences and independent errors with constant variance σ^2. Then, in P matched sets, the estimate of the average treatment effect has variance proportional to $\sigma^2(1 + 1/K)/P$. With the same number of matched sets—that is, with P fixed—increasing the number of controls, K, yields a less variable, more stable estimate, the constant of proportionality being $(1 + 1/K)$, which tends to 1 not to 0, as $K \to \infty$. See Hans K. Ury, "Efficiency of case-control studies with multiple controls per case: Continuous or dichotomous data," *Biometrics* (1975): 643–649.

32. Paul R. Rosenbaum, "Impact of multiple matched controls on design sensitivity in observational studies," *Biometrics* 69 (2013): 118–127. Magnitudes of impact on design sensitivity are given in table 4 of this article. For instance, in one Gaussian situation, using $K = 3$ controls rather than matched pairs, $K = 1$, increases the design sensitivity from $\tilde{\Gamma} = 3.3$ to $\tilde{\Gamma} = 4.1$, whereas using $K = 5$ controls raises the design sensitivity to $\tilde{\Gamma} = 4.6$. Additionally, using $K = 5$ controls rather than $K = 1$ control increases the sample size and reduces sampling variability. In the same situation, with 200 treated individuals at $\Gamma = 3$ in table 5 of this article, the combined effect of a larger design sensitivity plus a larger control group raises the power from 0.11 with $K = 1$ control to 0.76 with $K = 5$ controls, a substantial change.

33. I realize that this intuitive argument is neither convincing nor perfectly clear. Some technical arguments cannot be made perfectly clear without presenting technical detail; see Rosenbaum, "Impact of multiple matched controls," for that technical detail.

There are analogies that may be familiar to readers with technical background. The marginal distributions of two discrete random variables can limit the possible correlations between the random variables. In particular this is true of the permutation distributions of Z_i and r_{Ci} in a matched triple with two controls if the r_{Ci} take three distinct values. In other words, in a matched triple, the mere fact that the treatment assignments Z_i are binary and the r_{Ci} are not binary places an upper limit on their correlation.

For some related, albeit abstract, discussion, see Ward Whitt, "Bivariate distributions with given marginals," *Annals of Statistics* 4 (1976): 1280–1289; and Albert W. Marshall, "Copulas, marginals, and joint distributions," in *Distributions with Fixed Marginals and Related Topics,* ed. Ludger Rüschendorf, Berthold Schweizer, and Michael D. Taylor, 213–222 (Hayward, CA: Institute of Mathematical Statistics, 1996).

34. Samuel D. Pimentel, Frank Yoon, and Luke Keele, "Variable-ratio matching with fine balance in a study of the Peer Health Exchange," *Statistics in Medicine* 34 (2015): 4070–4082. In addition to being a published reference for Yoon's entire number, this article develops a new method of matching with a variable number of controls in combination with fine balance.

35. Kewei Ming and Paul R. Rosenbaum, "Substantial gains in bias reduction from matching with a variable number of controls," *Biometrics* 56 (2000): 118–124.

36. For an example of weighted boxplots in a control group with variable controls, see figures 3 and 4 and footnote 8 in Amelia Haviland, Daniel S. Nagin, and Paul R. Rosenbaum, "Combining propensity score matching and group-based trajectory analysis in an observational study," *Psychological Methods* 12 (2007): 247–267.

37. Paul R. Rosenbaum, "A characterization of optimal designs for observational studies," *Journal of the Royal Statistical Society,* Series B, 53 (1991): 597–610; Ben B. Hansen, "Full matching in an observational study of coaching for the SAT," *Journal of the American Statistical Association* 99 (2004): 609–618; Ben B. Hansen and Stephanie Olsen Klopfer, "Optimal full matching and related designs via network flows," *Journal of Computational and Graphical Statistics* 15 (2006): 609–627; Elizabeth A. Stuart and Kerry M. Green, "Using full matching to estimate causal effects in nonexperimental studies: examining the relationship between adolescent marijuana use and adult outcomes," *Developmental Psychology* 44 (2008): 395–406; Peter C. Austin and Elizabeth A. Stuart, "Optimal full matching for survival outcomes: A method that merits more widespread use," *Statistics in Medicine* 34 (2015): 3949–3967.

38. For a proof, see Rosenbaum, "A characterization of optimal designs," proposition 1 and corollary 3.

39. Wei Wu, Stephen G. West, and Jan N. Hughes, "Effect of grade retention in first grade on psychosocial outcomes," *Journal of Educational Psychology* 102 (2010): 135–152.

40. Jennifer L. Hill, Jane Waldfogel, Jeanne Brooks-Gunn, and Wen-Jui Han, "Maternal employment and child development: A fresh look using newer methods," *Developmental Psychology* 41 (2005): 833–850.

41. Phuong Nguyen-Hoang, "Fiscal effects of budget referendums: Evidence from New York school districts," *Public Choice* 150 (2012): 77–95.

42. Mitchell H. Gail, "Does cardiac transplantation prolong life? A reassessment," *Annals of Internal Medicine* 76 (1972): 815–817. Gail's critique affected the study of heart transplantation but also brought into focus a general problem. For further discussion of that problem, see the following publications: Nathan Mantel and David P. Byar, "Evaluation of response-time data involving transient states: An illustration using heart-transplant data," *Journal of the American Statistical Association* 69 (1974): 81–86; John Crowley and Marie Hu, "Covariance analysis of heart transplant survival data," *Journal of the American Statistical Association* 72 (1977): 27–36; Samy Suissa, "Immortal time bias in pharmacoepidemiology," *American Journal of Epidemiology* 167 (2008): 492–499; Mark J. Van der Laan and James M. Robins, *Unified Methods for Censored Longitudinal Data and Causality* (New York: Springer, 2003).

43. Yunfei Paul Li, Kathleen J. Propert, and Paul R. Rosenbaum, "Balanced risk-set matching," *Journal of the American Statistical Association* 96 (2001): 870–882; Bo Lu, "Propensity score matching with time-dependent covariates," *Biometrics* 61 (2005): 721–728. For two applications of risk-set matching, see the following works: Robert Apel, Arjan A. J. Blokland, Paul Nieuwbeerta, and Marieke van Schellen, "The impact of imprisonment on marriage and divorce: A risk-set matching approach," *Journal of Quantitative Criminology* 26 (2010): 269–300; Jeffrey H. Silber, Scott A. Lorch, Paul R. Rosenbaum, Barbara Medoff-Cooper, Susan Bakewell-Sachs, Andrea Millman, Lanyu Mi, Orit Even-Shoshan, and Gabriel J. Escobar, "Time to send the preemie home? Additional maturity at discharge and subsequent health care costs and outcomes," *Health Services Research* 44 (2009): 444–463.

44. Paul R. Rosenbaum, "The consequences of adjustment for a concomitant variable that has been affected by the treatment," *Journal of the Royal Statistical Society,* Series A, 147 (1984): 656–666.

45. José R. Zubizarreta, Mark Neuman, Jeffrey H. Silber, and Paul R. Rosenbaum, "Contrasting evidence within and between institutions that provide treatment in an observational study of alternate forms of anesthesia," *Journal of the American Statistical Association* 107 (2012): 901–915.

46. See Zubizarreta et al., "Contrasting evidence," for discussion of the algorithm and for the study's findings.

47. Jeffrey H. Silber, Paul R. Rosenbaum, Richard N. Ross, Justin M. Ludwig, Wei Wang, Bijan A. Niknam, Nabanita Mukherjee, Philip A. Saynisch, Orit Even-Shoshan, Rachel R. Kelz, and Lee A. Fleisher, "Template matching for auditing hospital cost and quality," *Health Services Research* 49 (2014): 1446–1474.

48. Silber et al., "Template matching for auditing," table 1.

49. Silber et al., "Template matching for auditing," table 2, and p. 1467.

12. Biases from General Dispositions

1. Lee J. Cronbach and Paul E. Meehl, "Construct validity in psychological tests," *Psychological Bulletin* 52 (1955): 281–302; Donald T. Campbell and Donald W. Fiske, "Convergent and discriminant validation by the multitrait-multimethod matrix," *Psychological Bulletin* 56 (1959): 81–105; Howard Wainer and Henry I. Braun, eds., *Test Validity* (Hillsdale, NJ: Lawrence Erlbaum, 1988); Samuel Messick, "Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning," *American Psychologist* 50 (1995): 741–749.

2. This chapter is a nontechnical exposition of the material in Paul R. Rosenbaum, "Differential effects and generic biases in observational studies," *Biometrika* 93 (2006): 573–586.

3. Georg Rasch, *Probabilistic Models for Some Intelligence and Attainment Tests* (Chicago: University of Chicago Press, 1960).

4. Write $Z_{ik} = 1$ if person i answers question k correctly, and $Z_{ik} = 0$ otherwise. There is a single latent variable u_i describing person i and a parameter ω_k describing the difficulty of question k. The Rasch model says,

$$\Pr(Z_{ik} = 1 \mid u_i) \; = \; \frac{\exp(u_i - \omega_k)}{1 + \exp(u_i - \omega_k)},$$

and the items are conditionally independent given the u_i.

5. Specifically, in Table 12.1, $Z_i = Z_{i1}$ and $Z_i^* = Z_{i2}$ and the parameters are $\exp(-\omega_1) = 1$ and $\exp(-\omega_2) = 3$. The high value of u_i is $\exp(u_i) = 3$, so $\Pr(Z_{i1} = 1 \mid u_i) = (3 \times 1)/(1 + 3 \times 1) = 3/4 = 0.75$, and $\Pr(Z_{i2} = 1 \mid u_i) = (3 \times 3)/(1 + 3 \times 3) = 9/100 = 0.90$, as in Table 12.1, and the chance that $\Pr(Z_{i1} = 1, Z_{i2} = 1 \mid u_i)$ is the product, $\Pr(Z_{i1} = 1 \mid u_i) \times \Pr(Z_{i2} = 1 \mid u_i) = 0.75 \times 0.90 = 0.675$, as in Table 12.1. The low value of u_i is $\exp(u_i) = 1$, so $\Pr(Z_{i1} = 1 \mid u_i) = (1 \times 1)/(1 + 1 \times 1) = 1/2 = .50$, and so on.

6. If the odds are 3-to-1, the probability is 0.75. Generally, in the Rasch model the probability $\Pr(Z_{ik} = 1 \mid u_i)$ that person i is exposed to treatment k depends on a quantity, u_i, which is not observed, so it differs from one person to the next and cannot be calcu-

lated from observed data. In contrast, if I look at a person who got either treatment $k=1$ or treatment $k=2$ but not both, $Z_{i1} + Z_{i2} = 1$, then for such a person, the chance that the person got treatment $k=1$ is

$$\Pr(Z_{i1} = 1 | u_i, Z_{i1} + Z_{i2} = 1) = \frac{\exp(u_i - \omega_1)}{\exp(u_i - \omega_1) + \exp(u_i - \omega_2)}$$

$$= \frac{\exp(\omega_2 - \omega_1)}{1 + \exp(\omega_2 - \omega_1)},$$

which no longer depends upon the unobserved u_i and is the same for every person i. Under the Rasch model, the differential comparison of $(Z_{i1}, Z_{i2}) = (1,0)$ and $(Z_{i1}, Z_{i2}) = (0,1)$ is unaffected by bias from u_i, where if the first treatment were studied alone, the comparison would be biased by u_i. This property is not restricted to the Rasch model and can hold for two items in an item response model without holding for all items; see Paul R. Rosenbaum, "Comparing item characteristic curves," *Psychometrika* 52 (1987): 217–233.

7. That is, under the Rasch model $Z_{i3} = z$ provides no additional information about Z_{i1} given $Z_{i1} + Z_{i2} = 1$, because

$$\Pr(Z_{i1} = 1 | u_i, Z_{i1} + Z_{i2} = 1, Z_{i3} = z) = \Pr(Z_{i1} = 1 | u_i, Z_{i1} + Z_{i2} = 1)$$

$$= \frac{\exp(\omega_2 - \omega_1)}{1 + \exp(\omega_2 - \omega_1)}.$$

See Tue Tjur, "A connection between Rasch's item analysis model and a multiplicative Poisson model," *Scandinavian Journal of Statistics* 9 (1982): 23–30, §4, table 4. We will see this phenomenon occur in an empirical example in Table 12.4.

8. Patrick L. McGeer, Michael Schulzer, and Edith G. McGeer, "Arthritis and anti-inflammatory agents as possible protective factors for Alzheimer's disease: A review of 17 epidemiologic studies," *Neurology* 47 (1996): 425–432; B. A. in 't Veld, L. J. Launer, M. M. B. Breteler, A. Hofman, and B. H. Ch. Stricker, "Pharmacologic agents associated with a preventive effect on Alzheimer's disease: A review of the epidemiologic evidence," *Epidemiologic Reviews* 24 (2002): 248–268.

9. in 't Veld et al., "Pharmacologic agents," 253.

10. Peter P. Zandi, James C. Anthony, Kathleen M. Hayden, Kala Mehta, Lawrence Mayer, and John C. S. Breitner, "Reduced incidence of AD with NSAID but not H2 receptor antagonists: The Cache County Study," *Neurology* 59 (2002): 880–886.

11. The use of differential effects in the study of drug safety is discussed by Robert D. Gibbons and Anup Amatya, *Statistical Methods for Drug Safety* (Boca Raton, FL: Chapman and Hall / CRC, 2016), §4.5.

12. *Fatal Accident Reporting System* (Washington, DC: National Highway Traffic Safety Administration, 2010).

13. Lawrence Evans, "The effectiveness of safety belts in preventing fatalities," *Accident Analysis and Prevention* 18 (1986): 229–241.

14. This is a form of interference between units in which the treatment given to one person affects someone else. See Chapter 5 for discussion of interference between units.

15. This analysis of 2010–2011 FARS data is a small part of analysis in Paul R. Rosenbaum, "Some counterclaims undermine themselves in observational studies," *Journal of the American Statistical Association* 110 (2015): 1389–1398.

16. Rosenbaum, "Some counterclaims undermine themselves," table 1.

17. *National Health and Nutrition Examination Survey* (Atlanta: U.S. Centers for Disease Control and Prevention), www.cdc.gov/nchs/nhanes.htm.

18. For additional detail about the example in this section, see Paul R. Rosenbaum, "Using differential comparisons in observational studies," *Chance* 26, no. 3 (2013): 18–25.

19. That is, the odds ratio is 6.0 with 95% confidence interval [4.0, 9.1].

20. This is a general property of monotone unidimensional latent variable models: any two items are non-negatively associated given any function of other items. Under such a model, no matter how many bad habits you control by matching, the next bad habit will nonetheless be more common among smokers than among nonsmokers. For specifics, see Paul W. Holland and Paul R. Rosenbaum, "Conditional association and unidimensionality in monotone latent variable models," *Annals of Statistics* 14 (1986): 1523–1543. To adequately adjust for a latent disposition to indulge bad habits, you need to overadjust for observed bad habits, and this is what differential effects do.

21. Here, person i has four potential responses, $(r_{11i}, r_{10i}, r_{01i}, r_{00i})$, under the four treatment combinations, $Z_i = z$ and $Z_i^* = z^*$ for $z = 0,1$ and $z^* = 0,1$. Then, by definition,

$$\pi_{z,z^*,i} = Pr(Z_i = z, Z_i^* = z^* | r_{11i}, r_{10i}, r_{01i}, r_{00i}, x_i, u_i)$$

Treatment assignment would be ignorable given the observed covariates x_i if $\pi_{z,z^*,i}$ does not depend upon $(r_{11i}, r_{10i}, r_{01i}, r_{00i}, u_i)$ although it may depend upon x_i. By definition, there are only generic biases if $\pi_{1,0,i}/\pi_{0,1,i}$ does not depend upon $(r_{11i}, r_{10i}, r_{01i}, r_{00i}, u_i)$ although it may depend upon x_i. The key point is that in the Rasch model and several other common models there is no generic bias even though treatment assignment is not ignorable; that is, $\pi_{z,z^*,i}$ depends upon u_i, but $\pi_{1,0,i}/\pi_{0,1,i}$ does not. In this case, the differential comparison is free of unmeasured bias, where the main effect of either treatment on its own would be biased by the failure to control for u_i. The definition of generic biases assumes much less than the Rasch model assumes; for instance, $\pi_{1,0,i}/\pi_{0,1,i}$ may vary with x_i.

22. If

$$\frac{\pi_{1,0,i}}{\pi_{0,1,i}} = \kappa(x_i),$$

then

$$\Pr \left(Z_i = 1, Z_i^* = 0 \mid r_{11i}, r_{10\,i}, r_{01i}, r_{00\,i}, x_i, u_i, Z_i + Z_i^* = 1\right) = \frac{\pi_{1,0,\,i}}{\pi_{1,0,\,i} + \pi_{0,1,\,i}}$$

is $\kappa(x_i) / \{1 + \kappa(x_i)\}$ and does not depend upon $(r_{11i}, r_{10i}, r_{01i}, r_{00i}, u_i)$.

23. The relevant mathematics occurs in §2 of José R. Zubizarreta, Dylan S. Small, and Paul R. Rosenbaum, "Isolation in the construction of natural experiments," *Annals of Applied Statistics* 8 (2014): 2096–2121.

24. Joshua D. Angrist and William N. Evans, "Children and their parent's labor supply: Evidence from exogenous variation in family size," *American Economic Review* 88 (1998): 450–477.

25. Zubizarreta, "Isolation in the construction of natural experiments." It is somewhat relevant that the data come from the 1980 U.S. Census and describe fertility in years leading up to that national census. Fertility treatment was much less common before 1980 than it is today. Fertility treatment increases the chance of twins. Twins were mostly luck before 1980, but today they can be unintended consequences of other deliberate choices.

13. Instruments

1. Traditionally, instruments were described in terms of structural equation models. The switch to speaking about instruments by analogy with randomized experimentation was encouraged and influenced by Joshua D. Angrist, Guido W. Imbens, and Donald B. Rubin, "Identification of causal effects using instrumental variables," *Journal of the American Statistical Association* 91 (1996): 444–455. One sometimes hears the term "instrumental variable" rather than "instrument." These two terms refer to the same thing: an instrumental variable records in a data file the activity of an instrument acting in the world. In addition to being less graceful, the term "instrumental variable" points to plans for using a data file, while "instrument" points to facts about the world. People who speak of an instrument as a thing in the world are more acutely aware that it is possible to err in describing the world; that is, one may mistakenly call something an instrument when it is not, thereby reaching mistaken scientific conclusions. In scientific work, being aware of the possibility of error is a good thing, and in this sense "instrument" is a better term than "instrumental variable."

2. Joseph P. Newhouse and Mark McClellan, "Econometrics in outcomes research: The use of instrumental variables," *Annual Review of Public Health* 19 (1998): 17–34. See also Michael Baiocchi, Jing Cheng, and Dylan S. Small, "Instrumental variable methods for causal inference," *Statistics in Medicine* 33 (2014): 2297–2340.

3. Fan Yang, José R. Zubizarreta, Dylan S. Small, Scott Lorch, and Paul R. Rosenbaum, "Dissonant conclusions when testing the validity of an instrumental variable," *American Statistician* 68 (2014): 253–263.

4. Mark McClellan, Barbara J. McNeil, and Joseph P. Newhouse, "Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Analysis using instrumental variables," *Journal of the American Medical Association* 272 (1994): 859–866; Scott A. Lorch, Michael Baiocchi, Corinne E. Ahlberg, and Dylan S. Small, "The differential impact of delivery hospital on the outcomes of premature infants," *Pediatrics* 130 (2012): 270–278; Mark D. Neuman, Paul R. Rosenbaum, Justin M. Ludwig, Jose R. Zubizarreta, and Jeffrey H. Silber, "Anesthesia technique, mortality, and length of stay after hip fracture surgery," *Journal of the American Medical Association* 311 (2014): 2508–2517.

5. M. Alan Brookhart, Philip Wang, Daniel H. Solomon, and Sebastian Schneeweiss, "Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable," *Epidemiology* 17 (2006): 268–275; Robert D. Gibbons and Anup Amatya, *Statistical Methods for Drug Safety* (Boca Raton, FL: Chapman and Hall / CRC, 2016), §4.4.

6. Joshua D. Angrist, "Lifetime earnings and the Vietnam era draft lottery: Evidence from social security administrative records," *American Economic Review* (1990): 313–336; Joshua D. Angrist, Stacey H. Chen, and Brigham R. Frandsen, "Did Vietnam veterans get sicker in the 1990s? The complicated effects of military service on self-reported health," *Journal of Public Economics* 94 (2010): 824–837; Joshua D. Angrist, Stacey H. Chen, and Jae Song, "Long-term consequences of Vietnam-era conscription: New estimates using social security data," *American Economic Review* 101 (2011): 334–338.

7. Paul W. Holland, "Causal inference, path analysis, and recursive structural equations models," *Sociological Methodology* 18 (1988): 449–484.

8. For details of such a test in the paired case, see proposition 2 in Mike Baiocchi, Dylan S. Small, Scott Lorch, and Paul R. Rosenbaum, "Building a stronger instrument in an observational study of perinatal care for premature infants," *Journal of the American Statistical Association* 105 (2010): 1285–1296.

9. More precisely, a $1 - \alpha$ confidence set for $\bar{\delta}/\bar{\eta}$ is obtained by testing at level α each possible value of ρ and retaining the values not rejected; then a $1 - \alpha$ confidence set is the shortest interval containing the confidence set. Such an interval may be infinitely long, say, (a,∞) or $(-\infty,a)$ or even $(-\infty,\infty)$, and such an infinite interval might indicate the data provide little information constraining the numerical value of $\bar{\delta}/\bar{\eta}$. In particular, infinite intervals are a convenient way to address the possibility that $\bar{\eta} = 0$. For related discussion, see Guido W. Imbens and Paul R. Rosenbaum, "Robust, accurate confidence intervals with a weak instrument: Quarter of birth and education," *Journal of the Royal Statistical Society, Series A,* 168 (2005): 109–126.

10. These two assumptions make a tidy story but they are not needed. The two assumptions play no role in inference about effect ratios. Additionally, an effect ratio can sometimes have a causal interpretation without these assumptions. For instance, if the effect of encouragement on lung function, $r_{Ti} - r_{Ci}$, was proportional to the effect on the number of cigarettes smoked, $s_{Ti} - s_{Ci}$, so that $r_{Ti} - r_{Ci} = \beta(s_{Ti} - s_{Ci})$, then the effect ratio is $\beta = \bar{\delta} / \bar{\eta}$. Here, β is a causal parameter, but it permits both reduced smoking rather than quitting and perverse effects of encouragement on smoking behavior, such as smoking more when encouraged to quit. For robust inference about β, see Imbens and Rosenbaum, "Robust, accurate confidence intervals."

11. Angrist et al., "Identification of causal effects."

12. This is one of the central results of Angrist et al., "Identification of causal effects," proposition 1.

13. Imbens and Rosenbaum, "Robust, accurate confidence intervals."

14. For example, the most popular method, two-stage least squares, can give crazy answers with weak instruments. A confidence interval gives the "wrong answer" if it breaks its promise. For instance, a 95% confidence interval promises to include the truth in 95% of studies, but if it actually includes the truth in only 40% of studies, then it broke its promise. Confidence intervals obtained from two-stage least squares break their promise in this sense when the instrument is weak. See John Bound, David A. Jaeger, and Regina M. Baker, "Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak," *Journal of the American Statistical Association* 90 (1995): 443–450.

15. Noncompliance is one thing, but loss to follow-up is another. Harry and Sally exhibit noncompliance by not receiving a full course of the assigned treatment. They were not lost to follow-up evaluation because the investigators continued to monitor their health outcomes. Whenever possible, patients who do not comply should be followed to determine their health outcomes. Noncompliance plus loss to follow-up evaluation is a much bigger problem than noncompliance alone.

16. For one instrumental estimate in a randomized trial with noncompliance, see Robert Greevy, Jeffrey H. Silber, Avital Cnaan, and Paul R. Rosenbaum, "Randomization inference with imperfect compliance in the ACE-inhibitor after anthracycline randomized trial," *Journal of the American Statistical Association* 99 (2004): 7–15.

17. Baiocchi et al., "Building a stronger instrument." The article by Baiocchi and colleagues is a methodological study with a clinical example. For a related clinical study, see Lorch et al., "Differential impact of delivery hospital."

18. Baiocchi et al., "Building a stronger instrument," table 1.

19. Baiocchi et al., "Building a stronger instrument," table 2.

20. Saying that encouragement is randomized is shorthand for: $0 < \Pr(Z=1 \mid x, r_T, r_C, s_T, s_C) = \Pr(Z=1 \mid x) < 1$, where Z is encouragement or distance, (r_T, r_C) is mortality, and (s_T, s_C) is the NICU level of the hospital of delivery.

21. Baiocchi et al., "Building a stronger instrument," tables 3 and 6.

22. More precisely, in a pair (i,j) discordant for infant mortality, a difference in (u_i,u_j) at most doubles the odds of living far from a high-level NICU and at most doubles the odds of death, in the absence of any effect on mortality. See Chapter 9 and the final columns in Table 9.1.

23. Chapter 10 asked a different but related question: Is it better to have fewer less heterogeneous pairs?

24. Baiocchi et al., "Building a stronger instrument," tables 3 and 6.

25. Dylan S. Small and Paul R. Rosenbaum, "War and wages: The strength of instrumental variables and their sensitivity to unobserved biases," *Journal of the American Statistical Association* 103 (2008): 924–933.

26. José R. Zubizarreta, Dylan S. Small, Neera K. Goyal, Scott Lorch, and Paul R. Rosenbaum, "Stronger instruments via integer programming in an observational study of late preterm birth outcomes," *Annals of Applied Statistics* 7 (2013): 25–50.

27. Yang et al., "Dissonant conclusions."

28. Baiocchi et al., "Building a stronger instrument"; Zubizarreta et al., "Stronger instruments"; Bo Lu, Robert Greevy, Xinyi Xu, and Cole Beck, "Optimal nonbipartite matching and its statistical applications," *American Statistician* 65 (2011): 21–30. In R, package nbpMatching implements optimal nonbipartite matching, as used in Baiocchi et al., "Building a stronger instrument."

14. Conclusion

1. See Table 5.6, where with 30 covariates, each at three levels, there are 2.1×10^{14} different types of people, far more types of people than there are people on Earth. You could not match treated and control individuals for 30 covariates, each with three levels, even if you had data on everyone who ever lived.

Appendix

1. Sir Ronald A. Fisher, *Design of Experiments* (Edinburgh: Oliver and Boyd, 1935). The small numerical example in my Chapter 3 with $I = 8$ patients is mathematically identical to the example in Chapter 2 of Fisher's *Design of Experiments,* although the surrounding narrative is quite different. For a concise survey of Fisher's contributions to statistics, see Leonard J. Savage, "On rereading R. A. Fisher," *Annals of Statistics* 4 (1976): 441–500. Essays about Fisher are collected in Stephen E. Fienberg and David V. Hinkely, *R. A. Fisher: An Appreciation* (New York: Springer-Verlag, 1980). For a biography of Fisher, see Joan Box, *R. A. Fisher, the Life of a Scientist* (New York: John Wiley and Sons, 1978). Fisher invented randomization as a statistical

method so that a feature of the design of the experiment justified, by a mathematical argument, a particular statistical inference. As an informal tool for equitable, blinded treatment assignment, randomization existed before Fisher; see Ian Hacking, "Origins of randomization in experimental design," *Isis* 79 (1988): 427–451.

2. David R. Cox and Nancy Reid, *The Theory of the Design of Experiments* (New York: Chapman and Hall / CRC, 2000). For the history of randomization in medical research, see Rosser Mathews, *Quantification and the Quest for Medical Certainty* (Princeton, NJ: Princeton University Press, 1995), 128.

3. Sonja M. McKinlay, "Experimentation in human populations," *Milbank Memorial Fund Quarterly: Health and Society* 59 (1981): 308–323; Paul Meier, "Statistics and medical experimentation," *Biometrics* 31 (1975): 511–529; Rosser Mathews, *Quantification and the Quest for Medical Certainty* (Princeton, NJ: Princeton University Press, 1995), 128; C. Frederick Mosteller and Robert F. Boruch, eds., *Evidence Matters: Randomized Trials in Education Research* (Washington, DC: Brookings Institution Press, 2002); David P. Farrington, "A short history of randomized experiments in criminology," *Evaluation Review* 27 (2003): 218–227; Richard A. Berk, "Randomized experiments as the bronze standard," *Journal of Experimental Criminology* 1 (2005): 417–433; Alan S. Gerber and Donald P. Green, *Field Experiments: Design, Analysis, and Interpretation* (New York: Norton, 2012); Steven D. Levitt and John A. List, "Field experiments in economics," *European Economic Review* 53 (2009): 1–18.

4. Jerzy Neyman, "On the application of probability theory to agricultural experiments: Essay on principles" (in Polish), *Roczniki Nauk Roiniczych Tom* 10 (1923): 1–51, reprinted in English in *Statistical Science* 5 (1990): 463–480.

5. B. L. Welch, "On the z-test in randomized blocks and Latin squares," *Biometrika* 42 (1937): 21–52; Oscar Kempthorne, *Design and Analysis of Experiments* (New York: John Wiley and Sons, 1952), chapter 8; Henry Scheffe, *The Analysis of Variance* (New York: John Wiley and Sons, 1959), chapter 9.

6. Donald B. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies," *Journal of Educational Psychology* 66 (1974): 688–701.

7. Ivan Oransky, "Obituary: Charles Frederick Mosteller," *Lancet* 368 (2006): 1062.

8. Nathan Mantel and William Haenszel, "Statistical aspects of retrospective studies of disease," *Journal of the National Cancer Institute* 22 (1957): 719–748; Nathan Mantel, "Chi-square tests with one degree of freedom; Extensions of the Mantel-Haenszel procedure," *Journal of the American Statistical Association* 58 (1963): 690–700; Nathan Mantel, "Synthetic retrospective studies and related topics," *Biometrics* 29 (1973): 479–486.

9. William G. Cochran, "The planning of observational studies of human populations (with discussion)," *Journal of the Royal Statistical Society,* Series A, 128 (1965): 134–155; William G. Cochran, "Reprint of 'Observational Studies' (with discussion),"

Observational Studies 1 (2015): 124–125, http://obsstudies.org/files/cochran_and _comments.pdf; William G. Cochran and Donald B. Rubin, "Controlling bias in observational studies: A review," *Sankhya,* Series A, 35 (1973): 417–446; Donald B. Rubin, "Assignment to treatment group on the basis of a covariate," *Journal of Educational Statistics* 2 (1977): 1–26; Donald B. Rubin, "Bayesian inference for causal effects: The role of randomization," *Annals of Statistics* 6 (1978): 34–58.

10. Paul R. Rosenbaum and Donald B. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika* 70 (1983): 41–55; Paul R. Rosenbaum and Donald B. Rubin, "Reducing bias in observational studies using subclassification on the propensity score," *Journal of the American Statistical Association* 79 (1984): 516–524; Paul R. Rosenbaum, "Conditional permutation tests and the propensity score in observational studies," *Journal of the American Statistical Association* 79 (1984): 565–574; Paul R. Rosenbaum and Donald B. Rubin, "Constructing a control group using multivariate matched sampling methods that incorporate the propensity score," *American Statistician* 39 (1985): 33–38; Paul R. Rosenbaum, "Model-based direct adjustment," *Journal of the American Statistical Association* 82 (1987): 387–394.

11. In epidemiology, see Jan P. Vandenbroucke, "When are observational studies as credible as randomized trials?," *Lancet* 363 (2004): 1728–1731; Jay S. Kaufman and Charles Poole, "Looking back on 'Causal Thinking in the Health Sciences,'" *Annual Review of Public Health* 21 (2000): 101–119. In economics, see Bruce D. Meyer, "Natural and quasi-experiments in economics," *Journal of Business and Economic Statistics* 12 (1995): 151–161; Joshua D. Angrist and Alan B. Krueger, "Empirical strategies in labor economics," *Handbook of Labor Economics* 3 (1999): 1277–1366; David S. Hamermesh, "The craft of labormetrics," *Industrial and Labor Relations Review* 53 (2000): 363–380; Mark R. Rosenzweig and Kenneth I. Wolpin, "Natural 'natural experiments' in economics," *Journal of Economic Literature* 38 (2000): 827–874. In psychology, see Donald T. Campbell, "Reforms as experiments," *American Psychologist* 24 (1969): 409–429; Michael Rutter, "Proceeding from observed correlation to causal inference: the use of natural experiments," *Perspectives in Psychological Science* 2 (2007): 377–395; Michael Rutter, "Natural experiments as a means of testing causal inferences," in *Causality: Statistical Perspectives and Applications,* ed. Carlo Berzuini, Philip Dawid, and Luisa Bernardinelli, 253–272 (New York: John Wiley, 2012). In political science, see Luke Keele and William Minozzi, "How much is Minnesota like Wisconsin? Assumptions and counterfactuals in causal inference with observational data," *Political Analysis* 21 (2013): 193–216; Luke Keele, "The statistics of causal inference: A view from political methodology," *Political Analysis* 23 (2015): 313–335; Jasjeet S. Sekhon and Rocio Titiunik, "When natural experiments are neither natural nor experiments," *American Political Science Review* 106 (2012): 35–57. For a textbook, see Thad Dunning, *Natural Experiments in the Social Sciences* (New York: Cambridge University Press, 2012).

12. Donald T. Campbell, "Factors relevant to the validity of experiments in social settings," *Psychological Bulletin* 54 (1957): 297–312; Donald L. Thistlethwaite and Donald T. Campbell, "Regression-discontinuity analysis: An alternative to the ex post facto experiment," *Journal of Educational Psychology* 51 (1960): 309–317; Donald T. Campbell and Julian C. Stanley, *Experimental and Quasi-experimental Designs for Research* (Chicago: Rand McNally, 1963); Donald T. Campbell, "Prospective: Artifact and control," in *Artifact in Behavioral Research,* ed. Robert Rosenthal and Ralph L. Rosnow, 264–286 (New York: Academic Press, 1969); Donald T. Campbell, "Reforms as experiments," *American Psychologist* 24 (1969): 409–429; William R. Shadish, Thomas D. Cook, and Donald T. Campbell, *Experimental and Quasi-experimental Designs for Generalized Causal Inference* (Boston: Houghton Mifflin, 2002).

13. Peter H. Rossi, Mark W. Lipsey, and Howard E. Freeman, *Evaluation* (Thousand Oaks, CA: Sage, 2004); Meyer, "Natural and quasi-experiments in economics"; Guido W. Imbens and Jeffrey M. Wooldridge, "Recent developments in the econometrics of program evaluation," *Journal of Economic Literature,* 47 (2009): 5–86; Michael R. Roberts and Toni M. Whited, "Endogeneity in empirical corporate finance," in *Handbook of the Economics of Finance,* vol. 2, part A, ed. George M. Constantinides, Milton Harris, and Rene M. Stulz, 493–572 (New York, Elsevier, 2013); Stephen G. West, Naihua Duan, Willo Pequegnat, Paul Gaist, Don C. Des Jarlais, David Holtgrave, Jose Szapocznik, Martin Fishbein, Bruce Rapkin, Michael Clatts, and Patricia D. Mullen, "Alternatives to the randomized controlled trial," *American Journal of Public Health* 98 (2008): 1359–1366.

14. Two interesting examples follow. Nada Eissa, "Taxation and labor supply of married women: The Tax Reform Act of 1986 as a natural experiment," *National Bureau of Economic Research Working Paper Series,* 1995, Working Paper No. 5023; Victor Lavy, "Effects of free choice among public schools," *Review of Economic Studies* 77 (2010): 1164–1191.

15. An introduction to evidence factors is contained in Paul R. Rosenbaum, "How to see more in observational studies: Some new quasi-experimental devices," *Annual Review of Statistics and Its Application* 2 (2015): 21–48.

16. Jerome Cornfield, William Haenszel, E. Cuyler Hammond, Abraham M. Lilienfeld, Michael B. Shimkin, and Ernst L, Wynder, "Smoking and lung cancer: Recent evidence and a discussion of some questions," *Journal of the National Cancer Institute* 22 (1959): 173–203; reprinted with discussion in *International Journal of Epidemiology* 38 (2009): 1175–1191.

17. Several R packages, available from `cran`, perform sensitivity analyses, including my `sensitivitymv` and `sensitivitymw`, Luke Keele's `rbounds`, Dylan Small's `SensitivityCaseControl`, and `sensitivity2x2xk`. Markus Gangl's `STATA` module `rbounds` performs sensitivity analyses. See also Paul R. Rosenbaum, "Two R packages for sensitivity analysis in observational studies," *Observa-*

tional Studies 1 (2015): 1–17. Several of these packages include data from observational studies and reproduce analyses from published articles.

18. Paul R. Rosenbaum, *Design of Observational Studies,* part III (New York: Springer, 2010); José R. Zubizarreta, Magdalena Cerdá, and Paul R. Rosenbaum, "Effect of the 2010 Chilean earthquake on posttraumatic stress: Reducing sensitivity to unmeasured bias through study design," *Epidemiology* 24 (2013): 79–87; Elizabeth A. Stuart and David B. Hanna, "Should epidemiologists be more sensitive to design sensitivity?," *Epidemiology* 24 (2013): 88–89; Paul R. Rosenbaum, "What aspects of the design of an observational study affect its sensitivity to bias from covariates that were not observed?," in *Looking Back: Proceedings of a Conference in Honor of Paul W. Holland,* ed. Neil J. Dorans and Sandip Sinharay, 87–114 (New York: Springer, 2011).

19. Noel S. Weiss, "Inferring causal relationships: elaboration of the criterion of 'dose-response,'" *American Journal of Epidemiology* 113 (1981): 487–490.

20. Elizabeth A. Stuart, "Matching methods for causal inference: A review and a look forward," *Statistical Science* 25 (2010): 1–21; Rosenbaum, *Design,* part II; Bo Lu, Robert Greevy, Xinyi Xu, and Cole Beck, "Optimal nonbipartite matching and its statistical applications," *American Statistician* 65 (2011): 21–30.

21. Ben Hansen's `optmatch` package in R constructs an optimal pair match or an optimal full match; see Ben B. Hansen, "Optmatch: Flexible, optimal matching for observational studies," *R News* 7 (2007): 18–24. Samuel Pimentel's package `rcbalance` in R performs optimal matching with various types of fine balance; see Samuel Pimentel, "Large sparse optimal matching with R package `rcbalance`," *Observational Studies* 2 (2016): 4–23. The `designmatch` package in R by Jose Zubizarreta and Cinar Kilcioglu offers several unique features made possible by a different set of algorithms for match optimization; see José R. Zubizarreta, "Using mixed integer programming for matching in an observational study of kidney failure after surgery," *Journal of the American Statistical Association* 107 (2012): 1360–1371. An R package `nbpMatching` by Bo Lu, Robert Greevy, Xinyi Xu, and Cole Beck implements optimal nonbipartite matching (see Lu et al., "Optimal nonbipartite matching").

22. Paul R. Rosenbaum, "Differential effects and generic biases in observational studies," *Biometrika* 93 (2006): 573–586; José R. Zubizarreta, Dylan S. Small, and Paul R. Rosenbaum, "Isolation in the construction of natural experiments," *Annals of Applied Statistics* 8 (2014): 2096–2121.

23. Joshua D. Angrist, Guido W. Imbens, and Donald B. Rubin, "Identification of causal effects using instrumental variables," *Journal of the American Statistical Association* 91 (1996): 444–455.

24. Mike Baiocchi, Dylan S. Small, Scott Lorch, and Paul R. Rosenbaum, "Building a stronger instrument in an observational study of perinatal care for premature infants," *Journal of the American Statistical Association* 105 (2010): 1285–1296.

GLOSSARY
Notation and Technical Terms

The glossary has two parts, one part for mathematical notation, the other for technical terms. The entry indicates the first chapter that uses the notation or technical term. If more than one chapter is mentioned, then the English word and the mathematical symbol are introduced in different chapters. The endnotes to particular chapters often provide mathematically precise definitions, most of which are not repeated in this glossary. **Bold** phrases denote cross-references within this glossary.

Notation

i, I Rather than refer to the people under study by name, we refer to them by an index number, i, from $i = 1, \ldots, I$. The I people in the study comprise a finite population. Chapter 2.

x_i The observed covariate for person i. Chapter 2.

u_i The unmeasured or unobserved covariate for person i. Chapter 2.

Z_i The treatment indicator for person i, with $Z_i = 1$ if person i received treatment or $Z_i = 0$ if person i received control. Chapter 2.

(r_{Ti}, r_{Ci}) The causal effect expressed, following Neyman and Rubin, as potential outcomes under treatment, r_{Ti}, and control, r_{Ci}. We see r_{Ti} if person i receives

treatment with $Z_i = 1$, or we see r_{Ci} if person i receives control with $Z_i = 0$, so we never see (r_{Ti}, r_{Ci}) jointly, thus making causal inference difficult. Chapter 2.

R_i The observed response of person i under the treatment actually received by person i. Here, if person i received treatment with $Z_i = 1$, then $R_i = r_{Ti}$, but if person i received control with $Z_i = 0$, then $R_i = r_{Ci}$. In either case, by simple algebra, $R_i = Z_i r_{Ti} + (1 - Z_i) r_{Ci}$. Chapter 2.

$\delta_i = r_{Ti} - r_{Ci}$ The causal effect difference, not seen for any person. Chapter 2.

π_i The probability that person i receives treatment, $\pi_i = \Pr(Z_i = 1)$. Implicitly, this probability makes use of the information in the observed covariates, x_i, the unobserved covariates, u_i, and the potential outcomes, (r_{Ti}, r_{Ci}). Chapter 2. For a precise definition see Chapter 5, note 33.

m The number of treated people, $m = Z_1 + \cdots + Z_I$. Chapter 2.

$\bar{v}$ An average or mean of a variable v_i over all I people, $i = 1, \ldots, I$ in the finite population under study. For instance, $\bar{r}_T$ is the average of r_{Ti} over all I people, and $\bar{r}_C$ is the average of r_{Ci} over all I people. The average treatment effect is $\bar{\delta} = \bar{r}_T - \bar{r}_C$ is the average of δ_i over all I people. Chapter 2.

$\hat{v}_T$ An average or mean of a variable v_i over the m treated people, with $Z_i = 1$. Chapter 2.

$\hat{v}_C$ An average or mean of a variable v_i over the $I - m$ control people, with $Z_i = 0$. Chapter 2.

T The upper left corner cell in Table 3.1, or equivalently the test statistic in Fisher's exact test for a 2×2 table recording treatment Z_i and observed binary outcome R_i. Because R_i is binary, $T = R_1 Z_1 + \cdots + R_I Z_I$. Chapter 3.

α The level of the test, conventionally $\alpha = 0.05$. A test has level α if the test rejects the null hypothesis when the P-value is less than or equal to α. A test that has level α will falsely reject a true null hypothesis H_0 with probability at most α. Chapter 3.

A The attributable effect, the total of δ_i over the m people in the treated group with $Z_i = 1$; that is, $A = \delta_1 Z_1 + \cdots + \delta_I Z_I$. Notice that $A = 0$ if Fisher's hypothesis of no effect is true. Chapter 3.

T^* The value of T in a uniformity trial, $T^* = r_{C1} Z_1 + \cdots + r_{CI} Z_I$. Equivalently, $A = T - T^*$. Chapter 3.

λ_x or $\lambda(x)$ The propensity score is the average λ_x of the treatment assignment probabilities π_i for all people i with a specific value x of the observed covariates x_i. Chapter 5. For a precise definition see Chapter 5, note 33.

Γ A parameter used to express the magnitude of departure from ignorable treatment assignment given x. The parameter is one of several essentially equivalent

measures of the bias in treatment assignment created by failure to adjust for an unmeasured covariate u. Two people, say, person i and person j, may differ in their odds of treatment by at most a factor of Γ; that is, $\pi_i / (1 - \pi_i)$ and $\pi_j / (1 - \pi_j)$ may differ by at most a factor of Γ. Equivalently, $1 / \Gamma \leq \{\pi_i(1 - \pi_j)\} / \{\pi_j(1 - \pi_i)\} \leq \Gamma$ for all i, j. See **Ignorable treatment assignment.** Chapter 9.

θ_p If pair p consists of person i and person j, one of whom was treated, $Z_i + Z_j = 1$, then θ_p is the probability that person i was treated rather than person j. Chapter 9. For a precise definition see Chapter 9, note 13.

(Λ, Δ) See **Amplification of a sensitivity analysis.** Chapter 9.

$\tilde{\Gamma}$ The design sensitivity. See **Design sensitivity.** Chapter 10.

Technical Terms

Amplification of a sensitivity analysis Restating the conclusions of a sensitivity analysis defined by one sensitivity parameter, say Γ, in terms of two or more other sensitivity parameters, say (Λ, Δ). An amplification permits a simple, compact sensitivity analysis to be performed and reported, but it permits detailed interpretation of that analysis without returning to the data for additional analysis. See **Sensitivity analysis.** Chapter 9.

Average treatment effect The average treatment effect, $\bar{\delta} = \bar{r}_T - \bar{r}_C$, is the average of the effect differences δ_i over all I people. It cannot be computed from data because we never see a causal effect, δ_i, but it can be estimated in a randomized experiment. Chapter 2.

Boxplot A convenient, compact picture of the distribution of data invented by John W. Tukey. See Figure 7.1 in Chapter 7.

Causal considerations A causal consideration is something you might reasonably consider in forming a judgement for which you are responsible. See the discussion of causal considerations as opposed to causal criteria in Chapter 10.

Causal effects See **Potential outcomes.**

Clustering A term borrowed from survey sampling. If a survey draws up a list of U.S. high schools, samples 50 high schools, then samples 20 students from each of the 50 sampled high schools, then the sample is said to be clustered. The schools are primary sampling units or clusters of students. The sampling of students is said to be two-stage sampling because students were sampled from sampled schools. The term clustering can be ambiguous or arbitrary when applied to administrative data in which two-stage sampling has not been used. See **Unit.** Chapters 5 and 10.

Confidence interval; confidence set   A 95% confidence set is a set calculated from data—hence, a random set because it depends on random data—having the property that it will cover the true value of a fixed parameter in at least 95% of studies. Equivalently, it is the set of values of the parameter not rejected by a 0.05-level test. A 95% confidence interval is the shortest interval containing a confidence set. **See** α**.**

Counterfactual   See **Potential Outcome.**

Covariate   A quantity measured before treatment assignment and hence unaffected by the treatment received subsequently. Chapters 1–2.

Covariate balance   If treated and control groups have the same distribution of covariates, there is covariate balance. Covariate balance is a property of the treated and control groups as groups. We often look to see if the observed covariates x are balanced. We may worry that an unobserved covariate u is not balanced. Chapters 1–2.

Design sensitivity, $\tilde{\Gamma}$. The limiting sensitivity to unmeasured bias as the sample size increases. The design sensitivity is a property of a sampling model and particular methods of analysis, evaluated when the sample size is large. Chapter 10.

Differential effects   The effect of giving one treatment in lieu of another, as distinct from no treatment. For instance, the effect of giving ibuprofen in lieu of acetaminophen, as distinct from the effect of giving ibuprofen instead of nothing. Some differential effects are immune to certain biases that affect comparisons of a treatment with no treatment. See **Generic unobserved bias.** Chapter 12.

Direct adjustment   An estimate of an average treatment effect formed by computing a separate estimate within each stratum defined by observed covariates and combining these separate estimates with suitable weights. Chapter 5.

Effect ratio   The ratio of two average treatment effects for different outcomes. Chapter 13.

Evidence factors   A study design contains two evidence factors if it permits two tests of the null hypothesis of no treatment effect that would be statistically independent if the null hypothesis were true and that are likely to be affected by very different biases from unmeasured covariates. Essentially, a study with two evidence factors provides an independent replicate of itself where the biases affecting the study and its replicate are very different. Chapter 7.

Exclusion restriction   See **Instrument.** Chapter 13.

Fine balance   In a matched comparison, a covariate is finely balanced if it has exactly the same distribution in treated and control groups without constraining

who is matched to whom. Related concepts are near-fine balance and refined balance. Chapter 11.

Fisher's exact test The randomization test for no treatment effect in a completely randomized experiment with a binary response. The data are described in a 2×2 table recording treatment Z_i and observed binary outcome R_i. Chapter 3.

Fisher's null hypothesis of no difference in treatment effects This null hypothesis, H_0, asserts that each person would have the same response under treatment as under control, $r_{Ti} - r_{Ci} = 0$ for $i = 1, \ldots, I$, or equivalently $\delta_i = 0$ for $i = 1, \ldots, I$. Often abbreviated as Fisher's hypothesis of no effect. Chapter 2.

Full matching In a full matching, each matched set contains either one treated individual and one or more controls, or else one control and one or more treated individuals. Chapter 11.

Generic unobserved bias An unobserved covariate u that promotes several treatments at the same time. Some generic biases can be eliminated or reduced using differential comparisons. See **Differential effects.** Chapter 12.

Ignorable treatment assignment Treatment assignment is ignorable if two people, say, person i and person j, with the same observed covariates, $x_i = x_j$, have the same probability of treatment, $\pi_i = \pi_j$. Chapter 5. For a precise definition see Chapter 5, note 33.

Instrument A random push that encourages a person to accept treatment, where the push can affect outcomes only to the extent that it alters the treatment received. The requirement that the push affects outcomes only by altering the treatment is called the exclusion restriction. A weak instrument is a gentle push, one that barely alters the treatment received. Chapter 13.

Interference between units The situation in which the treatment given to one person affects someone else. More precisely, the situation in which the treatment given to one unit can affect another unit. See **Unit.** Often confused with **Clustering.** Chapter 5.

Isolation Use of differential effects in conjunction with risk set matching. See **Differential effects.** See **Risk-set matching.** Chapter 12.

Mantel-Haenszel test A randomization test of no treatment effect with a binary response R and strata defined by observed covariates, x. Chapter 5.

McNemar's test The randomization test of no treatment effect in a paired randomized experiment with a binary outcome. Chapter 5.

Multiplicity Conventional statistical procedures, such as confidence intervals or hypothesis tests, come with a promise about the probability that they will err if used once. When these same procedures are used more than once, the

probability of error in at least one use gradually accumulates to a near-certainty of some error someplace. This is the problem of multiplicity, and various tools keep the problem under control. Chapter 8.

Natural experiment   The attempt to find in the world circumstances that resemble a randomized experiment. Chapter 6.

Null distribution   The distribution of a test statistic when the null hypothesis is true. Used in computing a *P*-value. Chapter 3.

Observational study   A study of the effects caused by treatments in which treatments are not randomly assigned to individuals, as they would be in a randomized experiment. Chapter 5.

Outcome   See **Potential Outcome.**

Potential outcome   Each individual i has a potential outcome under treatment, r_{Ti}, and a potential outcome under control, r_{Ci}. A causal effect is a comparison of what would have happened to person i under treatment, r_{Ti}, and under control, r_{Ci}. The central problem in causal inference is that we observed either r_{Ti} or r_{Ci} but never both. The use of potential outcomes to describe causal effects was introduced into statistics by Jerzy Neyman and Donald Rubin. Chapters 1–2.

Propensity score   The average λ_x of the treatment assignment probabilities π_i for all people i with a specific value x of the observed covariates x_i. Chapter 5. For a precise definition see Chapter 5, note 33.

P-value   A measure of the strength of evidence against a null hypothesis, H_0, evaluated using a test statistic, T. The *P*-value is a probability computed under the tentative assumption that the null hypothesis, H_0, is true. If the null hypothesis were true, the *P*-value would be the probability of a value of T as inconsistent or more inconsistent with the null hypothesis as the observed value of T. If the *P*-value is small, conventionally less than or equal to $\alpha = 0.05$, then T would have been unlikely to behave as it did were the null hypothesis true. In consequence, if the *P*-value is small, we have evidence against the truth of the null hypothesis. Chapter 3.

Quasi-experimental devices   Quasi-experimental devices reduce the ambiguity in an association between treatment and response by directing attention toward additional associations that might otherwise be ignored. Multiple control groups and counterparts are examples of quasi-experimental devices. Chapter 8.

Randomization. Randomized treatment assignment. Randomized experiment   Use of coins, dice or random numbers from the computer to assign people to treatments in a randomized experiment. Chapter 1.

Randomization on the basis of a covariate   A peculiar type of randomized experiment, proposed by Donald Rubin, in which the probability of treatment changes as a function of observed covariates. Chapter 5.

Randomization test   A test that uses the random assignment of treatments in a randomized experiment to test a null hypothesis, H_0, about treatment effects. A randomization test uses the actual random assignment of treatment and the null hypothesis, H_0, but it makes no additional modelling or sampling assumptions. Chapter 3.

Random sample   A random sample of size m from a finite population of size I draws m distinct people from the I people in such a way that every possible choice of m people has the same probability. More precisely, this is a simple random sample without replacement of size m from a finite population of size I. There are other, more complex types of random samples used, for example, by the U.S. Census Bureau. The treated group in a completely randomized experiment is a random sample of size m from the finite population of size I consisting of everyone in the experiment. Some people mistakenly use the term "random sample" to refer to any group of m people drawn from I people, even though they did not use computerized random numbers to draw a truly random sample. Chapters 2 and 3.

Risk-set matching   When people receive treatments at different times, risk-set matching pairs people who were similar up until the moment that one of them received treatment. Risk-set matching controls for the past, not for the future. Chapter 11.

Sensitivity analysis   Every application of mathematical ideas to worldly events depends upon assumptions. What if the assumptions are wrong? A sensitivity analysis relaxes the assumptions in a mathematical calculation, thereby displaying the degree to which the conclusions of that calculation change as the assumptions change. In this book, a sensitivity analysis relaxes the assumption of ignorable treatment assignment given observed covariates, thereby displaying the degree to which bias from an unobserved covariate might alter the conclusions of an observational study. Chapter 9.

Significance level   See *P-value.* Chapter 3.

Simple random sample   See **Random sample.**

Strata   Groups of similar people based on similar or identical values of observed covariates x_i.

Treatment effects   See **Potential outcomes.**

Uniformity trial   An experiment in which people are randomly assigned to treatment or control, but no treatment is applied, so treatment or control are merely

randomly assigned labels. Were Fisher's hypothesis of no effect true, the actual trial is essentially a uniformity trial. Chapter 3.

Unit A unit is an opportunity to apply or withhold the treatment. Mostly in this book a person is a unit. However, if one person can be given different treatments on different days, then a day for a person is a unit. If all people in the same school must receive the same treatment, then the school is the unit, not the student within the school. Chapter 5.

Weak instrument See **Instrument.** Chapter 13.

SUGGESTIONS FOR
FURTHER READING

Angrist, Joshua D., and Alan B. Krueger. "Empirical strategies in labor economics." *Handbook of Labor Economics* 3 (1999): 1277–1366.

Baiocchi, Michael, Jing Cheng, and Dylan S. Small. "Instrumental variable methods for causal inference." *Statistics in Medicine* 33 (2014): 2297–2340.

Cochran, William G. "The planning of observational studies of human populations (with discussion)." *Journal of the Royal Statistical Society,* Series A, 128 (1965): 134–155.

Cornfield, Jerome, William Haenszel, E. Cuyler Hammond, Abraham M. Lilienfeld, Michael B. Shimkin, and Ernst L. Wynder. "Smoking and lung cancer." Reprinted with discussion in the *International Journal of Epidemiology* 38 (2009): 1175–1191.

Cox, David R., and Nancy Reid. *The Theory of the Design of Experiments.* Boca Raton, FL: Chapman and Hall/CRC, 2000.

Hernan, Miguel, and James Robins. *Causal Inference.* Boca Raton, FL: Chapman and Hall/CRC, 2016.

Imbens, Guido W., and Donald B. Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences.* New York: Cambridge University Press, 2015.

Manski, Charles F. *Identification Problems in the Social Sciences.* Cambridge, MA: Harvard University Press, 1995.

Pearl, Judea. *Causality.* New York: Cambridge University Press, 2009.

Rosenbaum, Paul R. *Design of Observational Studies.* New York: Springer, 2010.

———. "How to see more in observational studies: Some new quasi-experimental devices." *Annual Review of Statistics and Its Application* 2 (2015): 21–48.

Rubin, Donald B. "Estimating causal effects of treatments in randomized and non-randomized studies." *Journal of Educational Psychology* 66 (1974): 688–701.

Rutter, Michael. *Identifying the Environmental Causes of Disease: How Should We Decide What to Believe and When to Take Action?* London: Academy of Medical Sciences, 2007.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. *Experimental and Quasi-experimental Designs for Generalized Causal Inference.* Boston: Houghton Mifflin, 2002.

Stuart, Elizabeth A. "Matching methods for causal inference: A review and a look forward." *Statistical Science* 25 (2010): 1–21.

Vandenbroucke, Jan P. "When are observational studies as credible as randomized trials?" *Lancet* 363 (2004): 1728–1731.

West, Stephen G., Naihua Duan, Willo Pequegnat, Paul Gaist, Don C. Des Jarlais, David Holtgrave, Jose Szapocznik, Martin Fishbein, Bruce Rapkin, Michael Clatts, and Patricia D. Mullen. "Alternatives to the randomized controlled trial." *American Journal of Public Health* 98 (2008): 1359–1366.

Zubizarreta, José R., Magdalena Cerdá, and Paul R. Rosenbaum. "Effect of the 2010 Chilean earthquake on posttraumatic stress: Reducing sensitivity to unmeasured bias through study design." *Epidemiology* 24 (2013): 79–87.

ACKNOWLEDGMENTS

INDEX