Chapter 1

Randomized Trials

KWAI CHANG CAINE: What happens in a man's life is already written. A man must move through life as his destiny wills.

OLD MAN: Yet each is free to live as he chooses. Though they seem opposite, both are true.

Kung Fu, Pilot

Our Path

Our path begins with experimental *random assignment,* both as a framework for causal questions and a benchmark by which the results from other methods are judged. We illustrate the awesome power of random assignment through two randomized evaluations of the effects of health insurance. The appendix to this chapter also uses the experimental framework to review the concepts and methods of statistical inference.

1.1 In Sickness and in Health (Insurance)

The Affordable Care Act (ACA) has proven to be one of the most controversial and interesting policy innovations we've seen. The ACA requires Americans to buy health insurance, with a tax penalty for those who don't voluntarily buy in. The question of the proper role of government in the market for health care has many angles. One is the causal effect of health insurance on health. The United States spends more of its GDP on health care than do other developed nations, yet Americans are surprisingly unhealthy. For example, Americans are more likely to be overweight and die sooner than their Canadian cousins, who spend only about two-thirds as much on care.

America is also unusual among developed countries in having no universal health insurance scheme. Perhaps there's a causal connection here.

Elderly Americans are covered by a federal program called Medicare, while some poor Americans (including most single mothers, their children, and many other poor children) are covered by Medicaid. Many of the working, prime-age poor, however, have long been uninsured. In fact, many uninsured Americans have chosen not to participate in an employer-provided insurance plan.[1] These workers, perhaps correctly, count on hospital emergency departments, which cannot turn them away, to address their health-care needs. But the emergency department might not be the best place to treat, say, the flu, or to manage chronic conditions like diabetes and hypertension that are so pervasive among poor Americans. The emergency department is not required to provide long-term care. It therefore stands to reason that government-mandated health insurance might yield a health dividend. The push for subsidized universal health insurance stems in part from the belief that it does.

The *ceteris paribus* question in this context contrasts the health of someone with insurance coverage to the health of the same person were they without insurance (other than an emergency department backstop). This contrast highlights a fundamental empirical conundrum: people are either insured or not. We don't get to see them both ways, at least not at the same time in exactly the same circumstances.

In his celebrated poem, "The Road Not Taken," Robert Frost used the metaphor of a crossroads to describe the causal effects of personal choice:

> 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;

[1] For more on this surprising fact, see Jonathan Gruber, "Covering the Uninsured in the United States," *Journal of Economic Literature*, vol. 46, no. 3, September 2008, pages 571–606.

Frost's traveler concludes:

> Two roads diverged in a wood, and I—
> I took the one less traveled by,
> And that has made all the difference.

The traveler claims his choice has mattered, but, being only one person, he can't be sure. A later trip or a report by other travelers won't nail it down for him, either. Our narrator might be older and wiser the second time around, while other travelers might have different experiences on the same road. So it is with any choice, including those related to health insurance: would uninsured men with heart disease be disease-free if they had insurance? In the novel *Light Years,* James Salter's irresolute narrator observes: "Acts demolish their alternatives, that is the paradox." We can't know what lies at the end of the road not taken.

We can't know, but evidence can be brought to bear on the question. This chapter takes you through some of the evidence related to paths involving health insurance. The starting point is the National Health Interview Survey (NHIS), an annual survey of the U.S. population with detailed information on health and health insurance. Among many other things, the NHIS asks: "Would you say your health in general is excellent, very good, good, fair, or poor?" We used this question to code an index that assigns 5 to excellent health and 1 to poor health in a sample of married 2009 NHIS respondents who may or may not be insured.[2] This index is our *outcome:* a measure we're interested in studying. The causal relation of interest here is determined by a variable that indicates coverage by private health insurance. We call this variable the *treatment,* borrowing from the literature on medical trials, although the treatments we're interested in need not be medical treatments like drugs or surgery. In this context, those with insurance can be thought of as the *treatment group;* those without insurance make up the *comparison* or *control group*. A good control group reveals the fate of the treated in a counterfactual world where they are not treated.

[2] Our sample is aged 26–59 and therefore does not yet qualify for Medicare.

The first row of Table 1.1 compares the average health index of insured and uninsured Americans, with statistics tabulated separately for husbands and wives.[3] Those with health insurance are indeed healthier than those without, a gap of about .3 in the index for men and .4 in the index for women. These are large differences when measured against the standard deviation of the health index, which is about 1. (Standard deviations, reported in square brackets in Table 1.1, measure variability in data. The chapter appendix reviews the relevant formula.) These large gaps might be the health dividend we're looking for.

Fruitless and Fruitful Comparisons

Simple comparisons, such as those at the top of Table 1.1, are often cited as evidence of causal effects. More often than not, however, such comparisons are misleading. Once again the problem is *other things equal,* or lack thereof. Comparisons of people with and without health insurance are not apples to apples; such contrasts are apples to oranges, or worse.

Among other differences, those with health insurance are better educated, have higher income, and are more likely to be working than the uninsured. This can be seen in panel B of Table 1.1, which reports the average characteristics of NHIS respondents who do and don't have health insurance. Many of the differences in the table are large (for example, a nearly 3-year schooling gap); most are statistically precise enough to rule out the hypothesis that these discrepancies are merely chance findings (see the chapter appendix for a refresher on statistical significance). It won't surprise you to learn that most variables tabulated here are highly correlated with health as well as with health insurance status. More-educated people, for example, tend to be healthier as well as being overrepresented in the insured group. This may be because more-educated people exercise more, smoke less, and are more likely to wear seat belts. It stands to reason that the difference in health between insured and uninsured NHIS

[3] An Empirical Notes section after the last chapter gives detailed notes for this table and most of the other tables and figures in the book.

TABLE 1.1
Health and demographic characteristics of insured and uninsured couples in the NHIS

	Husbands			Wives		
	Some HI (1)	No HI (2)	Difference (3)	Some HI (4)	No HI (5)	Difference (6)
			A. Health			
Health index	4.01 [.93]	3.70 [1.01]	.31 (.03)	4.02 [.92]	3.62 [1.01]	.39 (.04)
			B. Characteristics			
Nonwhite	.16	.17	−.01 (.01)	.15	.17	−.02 (.01)
Age	43.98	41.26	2.71 (.29)	42.24	39.62	2.62 (.30)
Education	14.31	11.56	2.74 (.10)	14.44	11.80	2.64 (.11)
Family size	3.50	3.98	−.47 (.05)	3.49	3.93	−.43 (.05)
Employed	.92	.85	.07 (.01)	.77	.56	.21 (.02)
Family income	106,467	45,656	60,810 (1,355)	106,212	46,385	59,828 (1,406)
Sample size	8,114	1,281		8,264	1,131	

Notes: This table reports average characteristics for insured and uninsured married couples in the 2009 National Health Interview Survey (NHIS). Columns (1), (2), (4), and (5) show average characteristics of the group of individuals specified by the column heading. Columns (3) and (6) report the difference between the average characteristic for individuals with and without health insurance (HI). Standard deviations are in brackets; standard errors are reported in parentheses.

respondents at least partly reflects the extra schooling of the insured.

Our effort to understand the causal connection between insurance and health is aided by fleshing out Frost's two-roads metaphor. We use the letter Y as shorthand for health, the outcome variable of interest. To make it clear when we're talking about specific people, we use subscripts as a stand-in for names: Y_i is the health of individual i. The outcome Y_i is recorded in our data. But, facing the choice of whether to pay for health insurance, person i has two *potential outcomes*, only one of which is observed. To distinguish one potential outcome from another, we add a second subscript: The road taken without health insurance leads to Y_{0i} (read this as "y-zero-i") for person i, while the road with health insurance leads to Y_{1i} (read this as "y-one–i") for person i. Potential outcomes lie at the end of each road one *might* take. The causal effect of insurance on health is the difference between them, written $Y_{1i} - Y_{0i}$.[4]

To nail this down further, consider the story of visiting Massachusetts Institute of Technology (MIT) student Khuzdar Khalat, recently arrived from Kazakhstan. Kazakhstan has a national health insurance system that covers all its citizens automatically (though you wouldn't go there just for the health insurance). Arriving in Cambridge, Massachusetts, Khuzdar is surprised to learn that MIT students must decide whether to opt in to the university's health insurance plan, for which MIT levies a hefty fee. Upon reflection, Khuzdar judges the MIT insurance worth paying for, since he fears upper respiratory infections in chilly New England. Let's say that $Y_{0i} = 3$ and $Y_{1i} = 4$ for $i =$ Khuzdar. For him, the causal effect of insurance is one step up on the NHIS scale:

$$Y_{1,\text{Khuzdar}} - Y_{0,\text{Khuzdar}} = 1.$$

Table 1.2 summarizes this information.

[4] Robert Frost's insights notwithstanding, econometrics isn't poetry. A modicum of mathematical notation allows us to describe and discuss subtle relationships precisely. We also use italics to introduce repeatedly used terms, such as *potential outcomes*, that have special meaning for masters of 'metrics.

TABLE 1.2
Outcomes and treatments for Khuzdar and Maria

	Khuzdar Khalat	Maria Moreño
Potential outcome without insurance: Y_{0i}	3	5
Potential outcome with insurance: Y_{1i}	4	5
Treatment (insurance status chosen): D_i	1	0
Actual health outcome: Y_i	4	5
Treatment effect: $Y_{1i} - Y_{0i}$	1	0

It's worth emphasizing that Table 1.2 is an imaginary table: some of the information it describes must remain hidden. Khuzdar will either buy insurance, revealing his value of Y_{1i}, or he won't, in which case his Y_{0i} is revealed. Khuzdar has walked many a long and dusty road in Kazakhstan, but even he cannot be sure what lies at the end of those not taken.

Maria Moreño is also coming to MIT this year; she hails from Chile's Andean highlands. Little concerned by Boston winters, hearty Maria is not the type to fall sick easily. She therefore passes up the MIT insurance, planning to use her money for travel instead. Because Maria has $Y_{0,\text{Maria}} = Y_{1,\text{Maria}} = 5$, the causal effect of insurance on her health is

$$Y_{1,\text{Maria}} - Y_{0,\text{Maria}} = 0.$$

Maria's numbers likewise appear in Table 1.2.

Since Khuzdar and Maria make different insurance choices, they offer an interesting comparison. Khuzdar's health is $Y_{\text{Khuzdar}} = Y_{1,\text{Khuzdar}} = 4$, while Maria's is $Y_{\text{Maria}} = Y_{0,\text{Maria}} = 5$. The difference between them is

$$Y_{\text{Khuzdar}} - Y_{\text{Maria}} = -1.$$

Taken at face value, this quantity—which we observe—suggests Khuzdar's decision to buy insurance is counterproductive. His MIT insurance coverage notwithstanding, insured Khuzdar's health is worse than uninsured Maria's.

In fact, the comparison between frail Khuzdar and hearty Maria tells us little about the causal effects of their choices. This can be seen by linking observed and potential outcomes as follows:

$$\begin{aligned} Y_{\text{Khuzdar}} - Y_{\text{Maria}} &= Y_{1,\text{Khuzdar}} - Y_{0,\text{Maria}} \\ &= \underbrace{Y_{1,\text{Khuzdar}} - Y_{0,\text{Khuzdar}}}_{1} + \underbrace{\{Y_{0,\text{Khuzdar}} - Y_{0,\text{Maria}}\}}_{-2}. \end{aligned}$$

The second line in this equation is derived by adding and subtracting $Y_{0,\text{Khuzdar}}$, thereby generating two hidden comparisons that determine the one we see. The first comparison, $Y_{1,\text{Khuzdar}} - Y_{0,\text{Khuzdar}}$, is the causal effect of health insurance on Khuzdar, which is equal to 1. The second, $Y_{0,\text{Khuzdar}} - Y_{0,\text{Maria}}$, is the difference between the two students' health status were both to decide against insurance. This term, equal to -2, reflects Khuzdar's relative frailty. In the context of our effort to uncover causal effects, the lack of comparability captured by the second term is called *selection bias*.

You might think that selection bias has something to do with our focus on particular individuals instead of on groups, where, perhaps, extraneous differences can be expected to "average out." But the difficult problem of selection bias carries over to comparisons of groups, though, instead of individual causal effects, our attention shifts to *average causal effects*. In a group of n people, average causal effects are written $Avg_n[Y_{1i} - Y_{0i}]$, where averaging is done in the usual way (that is, we sum individual outcomes and divide by n):

$$\begin{aligned} Avg_n[Y_{1i} - Y_{0i}] &= \frac{1}{n}\sum_{i=1}^{n}[Y_{1i} - Y_{0i}] \\ &= \frac{1}{n}\sum_{i=1}^{n} Y_{1i} - \frac{1}{n}\sum_{i=1}^{n} Y_{0i}. \end{aligned} \qquad (1.1)$$

The symbol $\sum_{i=1}^{n}$ indicates a sum over everyone from $i = 1$ to n, where n is the size of the group over which we are averaging. Note that both summations in equation (1.1) are taken over everybody in the group of interest. The average causal effect

of health insurance compares average health in hypothetical scenarios where everybody in the group does and does not have health insurance. As a computational matter, this is the average of individual causal effects like $Y_{1,\text{Khuzdar}} - Y_{0,\text{Khuzdar}}$ and $Y_{1,\text{Maria}} - Y_{0,\text{Maria}}$ for each student in our data.

An investigation of the average causal effect of insurance naturally begins by comparing the average health of groups of insured and uninsured people, as in Table 1.1. This comparison is facilitated by the construction of a *dummy variable*, D_i, which takes on the values 0 and 1 to indicate insurance status:

$$D_i = \begin{cases} 1 & \text{if } i \text{ is insured} \\ 0 & \text{otherwise.} \end{cases}$$

We can now write $Avg_n[Y_i|D_i = 1]$ for the average among the insured and $Avg_n[Y_i|D_i = 0]$ for the average among the uninsured. These quantities are averages *conditional* on insurance status.[5]

The average Y_i for the insured is necessarily an average of outcome Y_{1i}, but contains no information about Y_{0i}. Likewise, the average Y_i among the uninsured is an average of outcome Y_{0i}, but this average is devoid of information about the corresponding Y_{1i}. In other words, the road taken by those with insurance ends with Y_{1i}, while the road taken by those without insurance leads to Y_{0i}. This in turn leads to a simple but important conclusion about the difference in average health by insurance status:

$$\begin{aligned} &\textit{Difference in group means} \\ &\quad = Avg_n[Y_i|D_i = 1] - Avg_n[Y_i|D_i = 0] \\ &\quad = Avg_n[Y_{1i}|D_i = 1] - Avg_n[Y_{0i}|D_i = 0], \end{aligned} \tag{1.2}$$

[5] Order the n observations on Y_i so that the n_0 observations from the group indicated by $D_i = 0$ precede the n_1 observations from the $D_i = 1$ group. The conditional average

$$Avg_n[Y_i|D_i = 0] = \frac{1}{n_0}\sum_{i=1}^{n_0} Y_i$$

is the sample average for the n_0 observations in the $D_i = 0$ group. The term $Avg_n[Y_i|D_i = 1]$ is calculated analogously from the remaining n_1 observations.

an expression highlighting the fact that the comparisons in Table 1.1 tell us something about potential outcomes, though not necessarily what we want to know. We're after $Avg_n[Y_{1i} - Y_{0i}]$, an average causal effect involving everyone's Y_{1i} and everyone's Y_{0i}, but we see average Y_{1i} only for the insured and average Y_{0i} only for the uninsured.

To sharpen our understanding of equation (1.2), it helps to imagine that health insurance makes everyone healthier by a constant amount, κ. As is the custom among our people, we use Greek letters to label such *parameters*, so as to distinguish them from variables or data; this one is the letter "kappa." The *constant-effects assumption* allows us to write:

$$Y_{1i} = Y_{0i} + \kappa, \tag{1.3}$$

or, equivalently, $Y_{1i} - Y_{0i} = \kappa$. In other words, κ is both the individual and average causal effect of insurance on health. The question at hand is how comparisons such as those at the top of Table 1.1 relate to κ.

Using the constant-effects model (equation (1.3)) to substitute for $Avg_n[Y_{1i}|D_i = 1]$ in equation (1.2), we have:

$$\begin{aligned} &Avg_n[Y_{1i}|D_i = 1] - Avg_n[Y_{0i}|D_i = 0] \\ &\quad = \{\kappa + Avg_n[Y_{0i}|D_i = 1]\} - Avg_n[Y_{0i}|D_i = 0] \\ &\quad = \kappa + \big\{Avg_n[Y_{0i}|D_i = 1] - Avg_n[Y_{0i}|D_i = 0]\big\}. \end{aligned} \tag{1.4}$$

This equation reveals that health comparisons between those with and without insurance equal the causal effect of interest (κ) plus the difference in average Y_{0i} between the insured and the uninsured. As in the parable of Khuzdar and Maria, this second term describes selection bias. Specifically, the difference in average health by insurance status can be written:

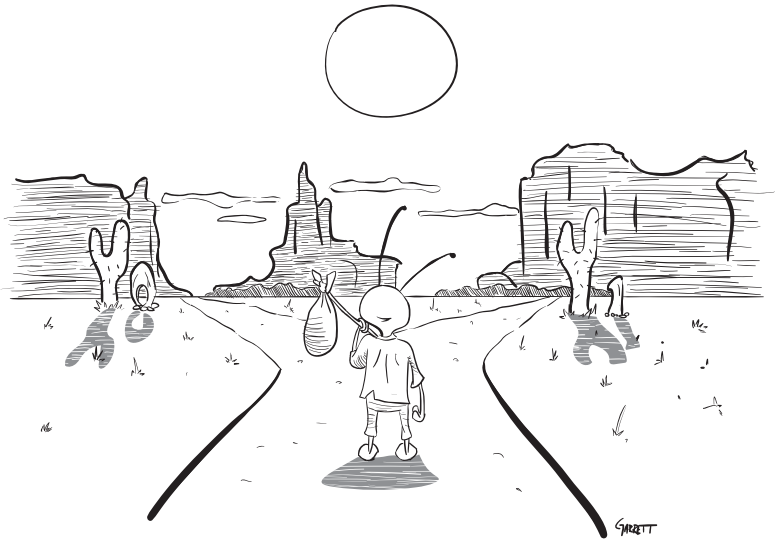
$$\begin{aligned} &\textit{Difference in group means} \\ &\quad = \textit{Average causal effect} + \textit{Selection bias}, \end{aligned}$$

where selection bias is defined as the difference in average Y_{0i} between the groups being compared.

How do we know that the difference in means by insurance status is contaminated by selection bias? We know because Y_{0i} is shorthand for everything about person i related to health, other than insurance status. The lower part of Table 1.1 documents important noninsurance differences between the insured and uninsured, showing that *ceteris* isn't *paribus* here in many ways. The insured in the NHIS are healthier for all sorts of reasons, including, perhaps, the causal effects of insurance. But the insured are also healthier because they are more educated, among other things. To see why this matters, imagine a world in which the causal effect of insurance is zero (that is, $\kappa = 0$). Even in such a world, we should expect insured NHIS respondents to be healthier, simply because they are more educated, richer, and so on.

We wrap up this discussion by pointing out the subtle role played by information like that reported in panel B of Table 1.1. This panel shows that the groups being compared differ in ways that we can observe. As we'll see in the next chapter, if the only source of selection bias is a set of differences in characteristics that we can observe and measure, selection bias is (relatively) easy to fix. Suppose, for example, that the only source of selection bias in the insurance comparison is education. This bias is eliminated by focusing on samples of people with the same schooling, say, college graduates. Education is the same for insured and uninsured people in such a sample, because it's the same for everyone in the sample.

The subtlety in Table 1.1 arises because when observed differences proliferate, so should our suspicions about unobserved differences. The fact that people with and without health insurance differ in many visible ways suggests that even were we to hold observed characteristics fixed, the uninsured would likely differ from the insured in ways we don't see (after all, the list of variables we can see is partly fortuitous). In other words, even in a sample consisting of insured and uninsured people with the same education, income, and employment status, the insured might have higher values of Y_{0i}. The principal challenge facing masters of 'metrics is elimination of the selection bias that arises from such unobserved differences.



Breaking the Deadlock: Just RANDomize

> My doctor gave me 6 months to live . . . but when I couldn't pay the bill, he gave me 6 months more.
>
> Walter Matthau

Experimental random assignment eliminates selection bias. The logistics of a randomized experiment, sometimes called a *randomized trial,* can be complex, but the logic is simple. To study the effects of health insurance in a randomized trial, we'd start with a sample of people who are currently uninsured. We'd then provide health insurance to a randomly chosen subset of this sample, and let the rest go to the emergency department if the need arises. Later, the health of the insured and uninsured groups can be compared. Random assignment makes this comparison *ceteris paribus*: groups insured and uninsured by random assignment differ only in their insurance status and any consequences that follow from it.

Suppose the MIT Health Service elects to forgo payment and tosses a coin to determine the insurance status of new students Ashish and Zandile (just this once, as a favor to their distinguished Economics Department). Zandile is insured if the toss comes up heads; otherwise, Ashish gets the coverage. A good start, but not good enough, since random assignment of two experimental subjects does not produce insured and uninsured apples. For one thing, Ashish is male and Zandile female. Women, as a rule, are healthier than men. If Zandile winds up healthier, it might be due to her good luck in having been born a woman and unrelated to her lucky draw in the insurance lottery. The problem here is that two is not enough to tango when it comes to random assignment. We must randomly assign treatment in a sample that's large enough to ensure that differences in individual characteristics like sex wash out.

Two randomly chosen groups, when large enough, are indeed comparable. This fact is due to a powerful statistical property known as the *Law of Large Numbers* (LLN). The LLN characterizes the behavior of sample averages in relation to sample size. Specifically, the LLN says that a sample average can be brought as close as we like to the average in the population from which it is drawn (say, the population of American college students) simply by enlarging the sample.

To see the LLN in action, play dice.[6] Specifically, roll a fair die once and save the result. Then roll again and average these two results. Keep on rolling and averaging. The numbers 1 to 6 are equally likely (that's why the die is said to be "fair"), so we can expect to see each value an equal number of times if we play long enough. Since there are six possibilities here, and all are equally likely, the expected outcome is an equally weighted average of each possibility, with weights equal to 1/6:

$$\begin{aligned}&\left(1 \times \tfrac{1}{6}\right) + \left(2 \times \tfrac{1}{6}\right) + \left(3 \times \tfrac{1}{6}\right) + \left(4 \times \tfrac{1}{6}\right) + \left(5 \times \tfrac{1}{6}\right) + \left(6 \times \tfrac{1}{6}\right) \\ &\quad = \frac{1+2+3+4+5+6}{6} = 3.5.\end{aligned}$$

[6] Six-sided cubes with one to six dots engraved on each side. There's an app for 'em on your smartphone.

This average value of 3.5 is called a *mathematical expectation;* in this case, it's the average value we'd get in infinitely many rolls of a fair die. The expectation concept is important to our work, so we define it formally here.

MATHEMATICAL EXPECTATION The mathematical expectation of a variable, Y_i, written $E[Y_i]$, is the population average of this variable. If Y_i is a variable generated by a random process, such as throwing a die, $E[Y_i]$ is the average in infinitely many repetitions of this process. If Y_i is a variable that comes from a sample survey, $E[Y_i]$ is the average obtained if everyone in the population from which the sample is drawn were to be enumerated.

Rolling a die only a few times, the average toss may be far from the corresponding mathematical expectation. Roll two times, for example, and you might get boxcars or snake eyes (two sixes or two ones). These average to values well away from the expected value of 3.5. But as the number of tosses goes up, the average across tosses reliably tends to 3.5. This is the LLN in action (and it's how casinos make a profit: in most gambling games, you can't beat the house in the long run, because the expected payout for players is negative). More remarkably, it needn't take too many rolls or too large a sample for a sample average to approach the expected value. The chapter appendix addresses the question of how the number of rolls or the size of a sample survey determines statistical accuracy.

In randomized trials, experimental samples are created by sampling from a population we'd like to study rather than by repeating a game, but the LLN works just the same. When sampled subjects are randomly divided (as if by a coin toss) into treatment and control groups, they come from the same underlying population. The LLN therefore promises that those in randomly assigned treatment and control samples will be similar if the samples are large enough. For example, we expect to see similar proportions of men and women in randomly assigned treatment and control groups. Random assignment also produces groups of about the same age and with similar

schooling levels. In fact, randomly assigned groups should be similar in every way, including in ways that we cannot easily measure or observe. This is the root of random assignment's awesome power to eliminate selection bias.

The power of random assignment can be described precisely using the following definition, which is closely related to the definition of mathematical expectation.

CONDITIONAL EXPECTATION The conditional expectation of a variable, Y_i, given a dummy variable, $D_i = 1$, is written $E[Y_i|D_i = 1]$. This is the average of Y_i in the population that has D_i equal to 1. Likewise, the conditional expectation of a variable, Y_i, given $D_i = 0$, written $E[Y_i|D_i = 0]$, is the average of Y_i in the population that has D_i equal to 0. If Y_i and D_i are variables generated by a random process, such as throwing a die under different circumstances, $E[Y_i|D_i = d]$ is the average of infinitely many repetitions of this process while holding the circumstances indicated by D_i fixed at d. If Y_i and D_i come from a sample survey, $E[Y_i|D_i = d]$ is the average computed when everyone in the population who has $D_i = d$ is sampled.

Because randomly assigned treatment and control groups come from the same underlying population, they are the same in every way, including their expected Y_{0i}. In other words, the conditional expectations, $E[Y_{0i}|D_i = 1]$ and $E[Y_{0i}|D_i = 0]$, are the same. This in turn means that:

RANDOM ASSIGNMENT ELIMINATES SELECTION BIAS When D_i is randomly assigned, $E[Y_{0i}|D_i = 1] = E[Y_{0i}|D_i = 0]$, and the difference in expectations by treatment status captures the causal effect of treatment:

$$\begin{aligned} &E[Y_i|D_i = 1] - E[Y_i|D_i = 0] \\ &\quad = E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0] \\ &\quad = E[Y_{0i} + \kappa|D_i = 1] - E[Y_{0i}|D_i = 0] \\ &\quad = \kappa + E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0] \\ &\quad = \kappa. \end{aligned}$$

Provided the sample at hand is large enough for the LLN to work its magic (so we can replace the conditional averages in equation (1.4) with conditional expectations), selection bias disappears in a randomized experiment. Random assignment works not by eliminating individual differences but rather by ensuring that the mix of individuals being compared is the same. Think of this as comparing barrels that include equal proportions of apples and oranges. As we explain in the chapters that follow, randomization isn't the only way to generate such *ceteris paribus* comparisons, but most masters believe it's the best.

When analyzing data from a randomized trial or any other research design, masters almost always begin with a check on whether treatment and control groups indeed look similar. This process, called *checking for balance*, amounts to a comparison of sample averages as in panel B of Table 1.1. The average characteristics in panel B appear dissimilar or unbalanced, underlining the fact that the data in this table don't come from anything like an experiment. It's worth checking for balance in this manner any time you find yourself estimating causal effects.

Random assignment of health insurance seems like a fanciful proposition. Yet health insurance coverage has twice been randomly assigned to large representative samples of Americans. The RAND Health Insurance Experiment (HIE), which ran from 1974 to 1982, was one of the most influential social experiments in research history. The HIE enrolled 3,958 people aged 14 to 61 from six areas of the country. The HIE sample excluded Medicare participants and most Medicaid and military health insurance subscribers. HIE participants were randomly assigned to one of 14 insurance plans. Participants did not have to pay insurance premiums, but the plans had a variety of provisions related to cost sharing, leading to large differences in the amount of insurance they offered.

The most generous HIE plan offered comprehensive care for free. At the other end of the insurance spectrum, three "catastrophic coverage" plans required families to pay 95% of their health-care costs, though these costs were capped as a propor-

tion of income (or capped at $1,000 per family, if that was lower). The catastrophic plans approximate a no-insurance condition. A second insurance scheme (the "individual deductible" plan) also required families to pay 95% of outpatient charges, but only up to $150 per person or $450 per family. A group of nine other plans had a variety of coinsurance provisions, requiring participants to cover anywhere from 25% to 50% of charges, but always capped at a proportion of income or $1,000, whichever was lower. Participating families enrolled in the experimental plans for 3 or 5 years and agreed to give up any earlier insurance coverage in return for a fixed monthly payment unrelated to their use of medical care.[7]

The HIE was motivated primarily by an interest in what economists call the price elasticity of demand for health care. Specifically, the RAND investigators wanted to know whether and by how much health-care use falls when the price of health care goes up. Families in the free care plan faced a price of zero, while coinsurance plans cut prices to 25% or 50% of costs incurred, and families in the catastrophic coverage and deductible plans paid something close to the sticker price for care, at least until they hit the spending cap. But the investigators also wanted to know whether more comprehensive and more generous health insurance coverage indeed leads to better health. The answer to the first question was a clear "yes": health-care consumption is highly responsive to the price of care. The answer to the second question is murkier.

Randomized Results

Randomized field experiments are more elaborate than a coin toss, sometimes regrettably so. The HIE was complicated by

[7] Our description of the HIE follows Robert H. Brook et al., "Does Free Care Improve Adults' Health? Results from a Randomized Controlled Trial," *New England Journal of Medicine*, vol. 309, no. 23, December 8, 1983, pages 1426–1434. See also Aviva Aron-Dine, Liran Einav, and Amy Finkelstein, "The RAND Health Insurance Experiment, Three Decades Later," *Journal of Economic Perspectives,* vol. 27, Winter 2013, pages 197–222, for a recent assessment.

having many small treatment groups, spread over more than a dozen insurance plans. The treatment groups associated with each plan are mostly too small for comparisons between them to be statistically meaningful. Most analyses of the HIE data therefore start by grouping subjects who were assigned to similar HIE plans together. We do that here as well.[8]

A natural grouping scheme combines plans by the amount of cost sharing they require. The three catastrophic coverage plans, with subscribers shouldering almost all of their medical expenses up to a fairly high cap, approximate a no-insurance state. The individual deductible plan provided more coverage, but only by reducing the cap on total expenses that plan participants were required to shoulder. The nine coinsurance plans provided more substantial coverage by splitting subscribers' health-care costs with the insurer, starting with the first dollar of costs incurred. Finally, the free plan constituted a radical intervention that might be expected to generate the largest increase in health-care usage and, perhaps, health. This categorization leads us to four groups of plans: catastrophic, deductible, coinsurance, and free, instead of the 14 original plans. The catastrophic plans provide the (approximate) no-

[8] Other HIE complications include the fact that instead of simply tossing a coin (or the computer equivalent), RAND investigators implemented a complex assignment scheme that potentially affects the statistical properties of the resulting analyses (for details, see Carl Morris, "A Finite Selection Model for Experimental Design of the Health Insurance Study," *Journal of Econometrics*, vol. 11, no. 1, September 1979, pages 43–61). Intentions here were good, in that the experimenters hoped to insure themselves against chance deviation from perfect balance across treatment groups. Most HIE analysts ignore the resulting statistical complications, though many probably join us in regretting this attempt to gild the random assignment lily. A more serious problem arises from the large number of HIE subjects who dropped out of the experiment and the large differences in attrition rates across treatment groups (fewer left the free plan, for example). As noted by Aron-Dine, Einav, and Finkelstein, "The RAND Experiment," *Journal of Economic Perspectives,* 2013, differential attrition may have compromised the experiment's validity. Today's "randomistas" do better on such nuts-and-bolts design issues (see, for example, the experiments described in Abhijit Banerjee and Esther Duflo, *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty,* Public Affairs, 2011).

insurance control, while the deductible, coinsurance, and free plans are characterized by increasing levels of coverage.

As with nonexperimental comparisons, a first step in our experimental analysis is to check for balance. Do subjects randomly assigned to treatment and control groups—in this case, to health insurance schemes ranging from little to complete coverage—indeed look similar? We gauge this by comparing demographic characteristics and health data collected before the experiment began. Because demographic characteristics are unchanging, while the health variables in question were measured before random assignment, we expect to see only small differences in these variables across the groups assigned to different plans.

In contrast with our comparison of NHIS respondents' characteristics by insurance status in Table 1.1, a comparison of characteristics across randomly assigned treatment groups in the RAND experiment shows the people assigned to different HIE plans to be similar. This can be seen in panel A of Table 1.3. Column (1) in this table reports averages for the catastrophic plan group, while the remaining columns compare the groups assigned more generous insurance coverage with the catastrophic control group. As a summary measure, column (5) compares a sample combining subjects in the deductible, coinsurance, and free plans with subjects in the catastrophic plans. Individuals assigned to the plans with more generous coverage are a little less likely to be female and a little less educated than those in the catastrophic plans. We also see some variation in income, but differences between plan groups are mostly small and are as likely to go one way as another. This pattern contrasts with the large and systematic demographic differences between insured and uninsured people seen in the NHIS data summarized in Table 1.1.

The small differences across groups seen in panel A of Table 1.3 seem likely to reflect chance variation that emerges naturally as part of the sampling process. In any statistical sample, chance differences arise because we're looking at one of many possible draws from the underlying population from which we've sampled. A new sample of similar size from the same population can be expected to produce comparisons that are similar—though not identical—to those in the table.

TABLE 1.3
Demographic characteristics and baseline health in the RAND HIE

	Means	Differences between plan groups			
	Catastrophic plan (1)	Deductible – catastrophic (2)	Coinsurance – catastrophic (3)	Free – catastrophic (4)	Any insurance – catastrophic (5)
A. Demographic characteristics					
Female	.560	−.023	−.025	−.038	−.030
		(.016)	(.015)	(.015)	(.013)
Nonwhite	.172	−.019	−.027	−.028	−.025
		(.027)	(.025)	(.025)	(.022)
Age	32.4	.56	.97	.43	.64
	[12.9]	(.68)	(.65)	(.61)	(.54)
Education	12.1	−.16	−.06	−.26	−.17
	[2.9]	(.19)	(.19)	(.18)	(.16)
Family income	31,603	−2,104	970	−976	−654
	[18,148]	(1,384)	(1,389)	(1,345)	(1,181)
Hospitalized last year	.115	.004	−.002	.001	.001
		(.016)	(.015)	(.015)	(.013)
B. Baseline health variables					
General health index	70.9	−1.44	.21	−1.31	−.93
	[14.9]	(.95)	(.92)	(.87)	(.77)
Cholesterol (mg/dl)	207	−1.42	−1.93	−5.25	−3.19
	[40]	(2.99)	(2.76)	(2.70)	(2.29)
Systolic blood pressure (mm Hg)	122	2.32	.91	1.12	1.39
	[17]	(1.15)	(1.08)	(1.01)	(.90)
Mental health index	73.8	−.12	1.19	.89	.71
	[14.3]	(.82)	(.81)	(.77)	(.68)
Number enrolled	759	881	1,022	1,295	3,198

Notes: This table describes the demographic characteristics and baseline health of subjects in the RAND Health Insurance Experiment (HIE). Column (1) shows the average for the group assigned catastrophic coverage. Columns (2)–(5) compare averages in the deductible, cost-sharing, free care, and any insurance groups with the average in column (1). Standard errors are reported in parentheses in columns (2)–(5); standard deviations are reported in brackets in column (1).

The question of how much variation we should expect from one sample to another is addressed by the tools of statistical inference.

The appendix to this chapter briefly explains how to quantify sampling variation with formal statistical tests. Such tests amount to the juxtaposition of differences in sample averages with their *standard errors*, the numbers in parentheses reported below the differences in averages listed in columns (2)–(5) of Table 1.3. The standard error of a difference in averages is a measure of its statistical precision: when a difference in sample averages is smaller than about two standard errors, the difference is typically judged to be a chance finding compatible with the hypothesis that the populations from which these samples were drawn are, in fact, the same.

Differences that are larger than about two standard errors are said to be *statistically significant:* in such cases, it is highly unlikely (though not impossible) that these differences arose purely by chance. Differences that are not statistically significant are probably due to the vagaries of the sampling process. The notion of statistical significance helps us interpret comparisons like those in Table 1.3. Not only are the differences in this table mostly small, only two (for proportion female in columns (4) and (5)) are more than twice as large as the associated standard errors. In tables with many comparisons, the presence of a few isolated statistically significant differences is usually also attributable to chance. We also take comfort from the fact that the standard errors in this table are not very big, indicating differences across groups are measured reasonably precisely.

Panel B of Table 1.3 complements the contrasts in panel A with evidence for reasonably good balance in *pre-treatment outcomes* across treatment groups. This panel shows no statistically significant differences in a pre-treatment index of general health. Likewise, pre-treatment cholesterol, blood pressure, and mental health appear largely unrelated to treatment assignment, with only a couple of contrasts close to statistical significance. In addition, although lower cholesterol in the

free group suggests somewhat better health than in the catastrophic group, differences in the general health index between these two groups go the other way (since lower index values indicate worse health). Lack of a consistent pattern reinforces the notion that these gaps are due to chance.

The first important finding to emerge from the HIE was that subjects assigned to more generous insurance plans used substantially more health care. This finding, which vindicates economists' view that demand for a good should go up when it gets cheaper, can be seen in panel A of Table 1.4.[9] As might be expected, hospital inpatient admissions were less sensitive to price than was outpatient care, probably because admissions decisions are usually made by doctors. On the other hand, assignment to the free care plan raised outpatient spending by two-thirds (169/248) relative to spending by those in catastrophic plans, while total medical expenses increased by 45%. These large gaps are economically important as well as statistically significant.

Subjects who didn't have to worry about the cost of health care clearly consumed quite a bit more of it. Did this extra care and expense make them healthier? Panel B in Table 1.4, which compares health indicators across HIE treatment groups, suggests not. Cholesterol levels, blood pressure, and summary indices of overall health and mental health are remarkably similar across groups (these outcomes were mostly measured 3 or 5 years after random assignment). Formal statistical tests show no statistically significant differences, as can be seen in the group-specific contrasts (reported in columns (2)–(4)) and in the differences in health between those in a catastrophic plan and everyone in the more generous insurance groups (reported in column (5)).

These HIE findings convinced many economists that generous health insurance can have unintended and undesirable

[9] The RAND results reported here are based on our own tabulations from the HIE public use file, as described in the Empirical Notes section at the end of the book. The original RAND results are summarized in Joseph P. Newhouse et al., *Free for All? Lessons from the RAND Health Insurance Experiment*, Harvard University Press, 1994.

TABLE 1.4
Health expenditure and health outcomes in the RAND HIE

	Means	Differences between plan groups			
	Catastrophic plan (1)	Deductible – catastrophic (2)	Coinsurance – catastrophic (3)	Free – catastrophic (4)	Any insurance – catastrophic (5)
		A. Health-care use			
Face-to-face visits	2.78 [5.50]	.19 (.25)	.48 (.24)	1.66 (.25)	.90 (.20)
Outpatient expenses	248 [488]	42 (21)	60 (21)	169 (20)	101 (17)
Hospital admissions	.099 [.379]	.016 (.011)	.002 (.011)	.029 (.010)	.017 (.009)
Inpatient expenses	388 [2,308]	72 (69)	93 (73)	116 (60)	97 (53)
Total expenses	636 [2,535]	114 (79)	152 (85)	285 (72)	198 (63)
		B. Health outcomes			
General health index	68.5 [15.9]	−.87 (.96)	.61 (.90)	−.78 (.87)	−.36 (.77)
Cholesterol (mg/dl)	203 [42]	.69 (2.57)	−2.31 (2.47)	−1.83 (2.39)	−1.32 (2.08)
Systolic blood pressure (mm Hg)	122 [19]	1.17 (1.06)	−1.39 (.99)	−.52 (.93)	−.36 (.85)
Mental health index	75.5 [14.8]	.45 (.91)	1.07 (.87)	.43 (.83)	.64 (.75)
Number enrolled	759	881	1,022	1,295	3,198

Notes: This table reports means and treatment effects for health expenditure and health outcomes in the RAND Health Insurance Experiment (HIE). Column (1) shows the average for the group assigned catastrophic coverage. Columns (2)–(5) compare averages in the deductible, cost-sharing, free care, and any insurance groups with the average in column (1). Standard errors are reported in parentheses in columns (2)–(5); standard deviations are reported in brackets in column (1).

consequences, increasing health-care usage and costs, without generating a dividend in the form of better health.[10]

1.2 The Oregon Trail

> Master Kan: Truth is hard to understand.
>
> Kwai Chang Caine: It is a fact, it is not the truth. Truth is often hidden, like a shadow in darkness.
>
> *Kung Fu,* Season 1, Episode 14

The HIE was an ambitious attempt to assess the impact of health insurance on health-care costs and health. And yet, as far as the contemporary debate over health insurance goes, the HIE might have missed the mark. For one thing, each HIE treatment group had at least catastrophic coverage, so financial liability for health-care costs was limited under every treatment. More importantly, today's uninsured Americans differ considerably from the HIE population: most of the uninsured are younger, less educated, poorer, and less likely to be working. The value of extra health care in such a group might be very different than for the middle class families that participated in the HIE.

One of the most controversial ideas in the contemporary health policy arena is the expansion of Medicaid to cover the currently uninsured (interestingly, on the eve of the RAND experiment, talk was of expanding Medicare, the public insurance program for America's elderly). Medicaid now covers families on welfare, some of the disabled, other poor children, and poor pregnant women. Suppose we were to expand Medicaid to cover those who don't qualify under current rules. How would such an expansion affect health-care spending? Would it shift treatment from costly and crowded emergency departments to possibly more effective primary care? Would Medicaid expansion improve health?

[10] Participants in the free plan had slightly better corrected vision than those in the other plans; see Brook et al., "Does Free Care Improve Health?" *New England Journal of Medicine*, 1983, for details.

Many American states have begun to "experiment" with Medicaid expansion in the sense that they've agreed to broaden eligibility, with the federal government footing most of the bill. Alas, these aren't real experiments, since everyone who is eligible for expanded Medicaid coverage gets it. The most convincing way to learn about the consequences of Medicaid expansion is to randomly offer Medicaid coverage to people in currently ineligible groups. Random assignment of Medicaid seems too much to hope for. Yet, in an awesome social experiment, the state of Oregon recently offered Medicaid to thousands of randomly chosen people in a publicly announced health insurance lottery.

We can think of Oregon's health insurance lottery as randomly selecting winners and losers from a pool of registrants, though coverage was not automatic, even for lottery winners. Winners won the opportunity to apply for the state-run Oregon Health Plan (OHP), the Oregon version of Medicaid. The state then reviewed these applications, awarding coverage to Oregon residents who were U.S. citizens or legal immigrants aged 19–64, not otherwise eligible for Medicaid, uninsured for at least 6 months, with income below the federal poverty level, and few financial assets. To initiate coverage, lottery winners had to document their poverty status and submit the required paperwork within 45 days.

The rationale for the 2008 OHP lottery was fairness and not research, but it's no less awesome for that. The Oregon health insurance lottery provides some of the best evidence we can hope to find on the costs and benefits of insurance coverage for the currently uninsured, a fact that motivated research on OHP by MIT master Amy Finkelstein and her coauthors.[11]

[11] See Amy Finkelstein et al., "The Oregon Health Insurance Experiment: Evidence from the First Year," *Quarterly Journal of Economics,* vol. 127, no. 3, August 2012, pages 1057–1106; Katherine Baicker et al., "The Oregon Experiment—Effects of Medicaid on Clinical Outcomes," *New England Journal of Medicine*, vol. 368, no. 18, May 2, 2013, pages 1713–1722; and Sarah Taubman et al., "Medicaid Increases Emergency Department Use: Evidence from Oregon's Health Insurance Experiment," *Science*, vol. 343, no. 6168, January 17, 2014, pages 263–268.

Roughly 75,000 lottery applicants registered for expanded coverage through the OHP. Of these, almost 30,000 were randomly selected and invited to apply for OHP; these winners constitute the OHP treatment group. The other 45,000 constitute the OHP control sample.

The first question that arises in this context is whether OHP lottery winners were more likely to end up insured as a result of winning. This question is motivated by the fact that some applicants qualified for regular Medicaid even without the lottery. Panel A of Table 1.5 shows that about 14% of controls (lottery losers) were covered by Medicaid in the year following the first OHP lottery. At the same time, the second column, which reports differences between the treatment and control groups, shows that the probability of Medicaid coverage increased by 26 percentage points for lottery winners. Column (4) shows a similar increase for the subsample living in and around Portland, Oregon's largest city. The upshot is that OHP lottery winners were insured at much higher rates than were lottery losers, a difference that might have affected their use of health care and their health.[12]

The OHP treatment group (that is, lottery winners) used more health-care services than they otherwise would have. This can also be seen in Table 1.5, which shows estimates of changes in service use in the rows below the estimate of the OHP effect on Medicaid coverage. The hospitalization rate increased by about half a percentage point, a modest though statistically significant effect. Emergency department visits, outpatient visits, and prescription drug use all increased markedly. The fact that the number of emergency department visits rose about 10%, a precisely estimated effect (the standard error associated with this estimate, reported in column (4), is .029), is especially noteworthy. Many policymakers hoped and expected health insurance to shift formerly uninsured patients

[12] Why weren't all OHP lottery winners insured? Some failed to submit the required paperwork on time, while about half of those who did complete the necessary forms in a timely fashion turned out to be ineligible on further review.

Table 1.5
OHP effects on insurance coverage and health-care use

	Oregon		Portland area	
Outcome	Control mean (1)	Treatment effect (2)	Control mean (3)	Treatment effect (4)
A. Administrative data				
Ever on Medicaid	.141	.256 (.004)	.151	.247 (.006)
Any hospital admissions	.067	.005 (.002)		
Any emergency department visit			.345	.017 (.006)
Number of emergency department visits			1.02	.101 (.029)
Sample size	74,922		24,646	
B. Survey data				
Outpatient visits (in the past 6 months)	1.91	.314 (.054)		
Any prescriptions?	.637	.025 (.008)		
Sample size	23,741			

Notes: This table reports estimates of the effect of winning the Oregon Health Plan (OHP) lottery on insurance coverage and use of health care. Odd-numbered columns show control group averages. Even-numbered columns report the regression coefficient on a dummy for lottery winners. Standard errors are reported in parentheses.

away from hospital emergency departments toward less costly sources of care.

Finally, the proof of the health insurance pudding appears in Table 1.6: lottery winners in the statewide sample report a modest improvement in the probability they assess their health as being good or better (an effect of .039, which can be

TABLE 1.6
OHP effects on health indicators and financial health

	Oregon		Portland area	
Outcome	Control mean (1)	Treatment effect (2)	Control mean (3)	Treatment effect (4)
A. Health indicators				
Health is good	.548	.039 (.008)		
Physical health index			45.5	.29 (.21)
Mental health index			44.4	.47 (.24)
Cholesterol			204	.53 (.69)
Systolic blood pressure (mm Hg)			119	−.13 (.30)
B. Financial health				
Medical expenditures >30% of income			.055	−.011 (.005)
Any medical debt?			.568	−.032 (.010)
Sample size	23,741		12,229	

Notes: This table reports estimates of the effect of winning the Oregon Health Plan (OHP) lottery on health indicators and financial health. Odd-numbered columns show control group averages. Even-numbered columns report the regression coefficient on a dummy for lottery winners. Standard errors are reported in parentheses.

compared with a control mean of .55; the Health is Good variable is a dummy). Results from in-person interviews conducted in Portland suggest these gains stem more from improved mental rather than physical health, as can be seen in the second and third rows in column (4) (the health variables in the Portland sample are indices ranging from 0 to 100). As in the RAND

experiment, results from Portland suggest physical health indicators like cholesterol and blood pressure were largely unchanged by increased access to OHP insurance.

The weak health effects of the OHP lottery disappointed policymakers who looked to publicly provided insurance to generate a health dividend for low-income Americans. The fact that health insurance increased rather than decreased expensive emergency department use is especially frustrating. At the same time, panel B of Table 1.6 reveals that health insurance provided the sort of financial safety net for which it was designed. Specifically, households winning the lottery were less likely to have incurred large medical expenses or to have accumulated debt generated by the need to pay for health care. It may be this improvement in financial health that accounts for improved mental health in the treatment group.

It also bears emphasizing that the financial and health effects seen in Table 1.6 most likely come from the 25% of the sample who obtained insurance as a result of the lottery. Adjusting for the fact that insurance status was unchanged for many winners shows that gains in financial security and mental health for the one-quarter of applicants who were insured as a result of the lottery were considerably larger than simple comparisons of winners and losers would suggest. Chapter 3, on instrumental variables methods, details the nature of such adjustments. As you'll soon see, the appropriate adjustment here amounts to the division of win/loss differences in outcomes by win/loss differences in the probability of insurance. This implies that the effect of being insured is as much as four times larger than the effect of winning the OHP lottery (statistical significance is unchanged by this adjustment).

The RAND and Oregon findings are remarkably similar. Two ambitious experiments targeting substantially different populations show that the use of health-care services increases sharply in response to insurance coverage, while neither experiment reveals much of an insurance effect on physical health. In 2008, OHP lottery winners enjoyed small but noticeable improvements in mental health. Importantly, and not coincidentally, OHP also succeeded in insulating many lottery winners from the financial consequences of poor health, just as a

good insurance policy should. At the same time, these studies suggest that subsidized public health insurance should not be expected to yield a dramatic health dividend.

MASTER JOSHWAY: In a nutshell, please, Grasshopper.

GRASSHOPPER: Causal inference compares potential outcomes, descriptions of the world when alternative roads are taken.

MASTER JOSHWAY: Do we compare those who took one road with those who took another?

GRASSHOPPER: Such comparisons are often contaminated by selection bias, that is, differences between treated and control subjects that exist even in the absence of a treatment effect.

MASTER JOSHWAY: Can selection bias be eliminated?

GRASSHOPPER: Random assignment to treatment and control conditions eliminates selection bias. Yet even in randomized trials, we check for balance.

MASTER JOSHWAY: Is there a single causal truth, which all randomized investigations are sure to reveal?

GRASSHOPPER: I see now that there can be many truths, Master, some compatible, some in contradiction. We therefore take special note when findings from two or more experiments are similar.

Masters of 'Metrics: From Daniel to R. A. Fisher

The value of a control group was revealed in the Old Testament. The Book of Daniel recounts how Babylonian King Nebuchadnezzar decided to groom Daniel and other Israelite captives for his royal service. As slavery goes, this wasn't a bad gig, since the king ordered his captives be fed "food and wine from the king's table." Daniel was uneasy about the rich diet, however, preferring modest vegetarian fare. The king's chamberlains initially refused Daniel's special meals request, fearing that his diet would prove inadequate for one called on

to serve the king. Daniel, not without chutzpah, proposed a controlled experiment: "Test your servants for ten days. Give us nothing but vegetables to eat and water to drink. Then compare our appearance with that of the young men who eat the royal food, and treat your servants in accordance with what you see" (Daniel 1, 12–13). The Bible recounts how this experiment supported Daniel's conjecture regarding the relative healthfulness of a vegetarian diet, though as far as we know Daniel himself didn't get an academic paper out of it.

Nutrition is a recurring theme in the quest for balance. Scurvy, a debilitating disease caused by vitamin C deficiency, was the scourge of the British Navy. In 1742, James Lind, a surgeon on HMS *Salisbury,* experimented with a cure for scurvy. Lind chose 12 seamen with scurvy and started them on an identical diet. He then formed six pairs and treated each of the pairs with a different supplement to their daily food ration. One of the additions was an extra two oranges and one lemon (Lind believed an acidic diet might cure scurvy). Though Lind did not use random assignment, and his sample was small by our standards, he was a pioneer in that he chose his 12 study members so they were "as similar as I could have them." The citrus eaters—Britain's first limeys—were quickly and incontrovertibly cured, a life-changing empirical finding that emerged from Lind's data even though his theory was wrong.[13]

Almost 150 years passed between Lind and the first recorded use of experimental random assignment. This was by Charles Peirce, an American philosopher and scientist, who experimented with subjects' ability to detect small differences in weight. In a less-than-fascinating but methodologically significant 1885 publication, Peirce and his student Joseph Jastrow explained how they varied experimental conditions according to draws from a pile of playing cards.[14]

[13] Lind's experiment is described in Duncan P. Thomas, "Sailors, Scurvy, and Science," *Journal of the Royal Society of Medicine,* vol. 90, no. 1, January 1997, pages 50–54.

[14] Charles S. Peirce and Joseph Jastrow, "On Small Differences in Sensation," *Memoirs of the National Academy of Sciences,* vol. 3, 1885, pages 75–83.

The idea of a randomized controlled trial emerged in earnest only at the beginning of the twentieth century, in the work of statistician and geneticist Sir Ronald Aylmer Fisher, who analyzed data from agricultural experiments. Experimental random assignment features in Fisher's 1925 *Statistical Methods for Research Workers* and is detailed in his landmark *The Design of Experiments,* published in 1935.[15]

Fisher had many fantastically good ideas and a few bad ones. In addition to explaining the value of random assignment, he invented the statistical method of maximum likelihood. Along with 'metrics master Sewall Wright (and J.B.S. Haldane), he launched the field of theoretical population genetics. But he was also a committed eugenicist and a proponent of forced sterilization (as was regression master Sir Francis Galton, who coined the term "eugenics"). Fisher, a lifelong pipe smoker, was

[15] Ronald A. Fisher, *Statistical Methods for Research Workers,* Oliver and Boyd, 1925, and Ronald A. Fisher, *The Design of Experiments*, Oliver and Boyd, 1935.

also on the wrong side of the debate over smoking and health, due in part to his strongly held belief that smoking and lung cancer share a common genetic origin. The negative effect of smoking on health now seems well established, though Fisher was right to worry about selection bias in health research. Many lifestyle choices, such as low-fat diets and vitamins, have been shown to be unrelated to health outcomes when evaluated with random assignment.

Appendix: Mastering Inference

YOUNG CAINE: I am puzzled.

MASTER PO: That is the beginning of wisdom.
Kung Fu, Season 2, Episode 25

This is the first of a number of appendices that fill in key econometric and statistical details. You can spend your life studying statistical inference; many masters do. Here we offer a brief sketch of essential ideas and basic statistical tools, enough to understand tables like those in this chapter.

The HIE is based on a sample of participants drawn (more or less) at random from the population eligible for the experiment. Drawing another sample from the same population, we'd get somewhat different results, but the general picture should be similar if the sample is large enough for the LLN to kick in. How can we decide whether statistical results constitute strong evidence or merely a lucky draw, unlikely to be replicated in repeated samples? How much sampling variance should we expect? The tools of formal statistical inference answer these questions. These tools work for all of the econometric strategies of concern to us. Quantifying sampling uncertainty is a necessary step in any empirical project and on the road to understanding statistical claims made by others. We explain the basic inference idea here in the context of HIE treatment effects.

The task at hand is the quantification of the uncertainty associated with a particular sample average and, especially,

groups of averages and the differences among them. For example, we'd like to know if the large differences in health-care expenditure across HIE treatment groups can be discounted as a chance finding. The HIE samples were drawn from a much larger data set that we think of as covering the population of interest. The HIE population consists of all families eligible for the experiment (too young for Medicare and so on). Instead of studying the many millions of such families, a much smaller group of about 2,000 families (containing about 4,000 people) was selected at random and then randomly allocated to one of 14 plans or treatment groups. Note that there are two sorts of randomness at work here: the first pertains to the construction of the study sample and the second to how treatment was allocated to those who were sampled. *Random sampling* and *random assignment* are closely related but distinct ideas.

A World without Bias

We first quantify the uncertainty induced by random sampling, beginning with a single sample average, say, the average health of everyone in the sample at hand, as measured by a health index. Our target is the corresponding population average health index, that is, the mean over everyone in the population of interest. As we noted on p. 14, the population mean of a variable is called its *mathematical expectation,* or just *expectation* for short. For the expectation of a variable, Y_i, we write $E[Y_i]$. Expectation is intimately related to formal notions of *probability.* Expectations can be written as a weighted average of all possible values that the variable Y_i can take on, with weights given by the probability these values appear in the population. In our dice-throwing example, these weights are equal and given by 1/6 (see Section 1.1).

Unlike our notation for averages, the symbol for expectation does not reference the sample size. That's because expectations are population quantities, defined without reference to a particular sample of individuals. For a given population, there is only one $E[Y_i]$, while there are many $Avg_n[Y_i]$, depending on how we choose n and just who ends up in our sample. Because

$E[Y_i]$ is a fixed feature of a particular population, we call it a *parameter*. Quantities that vary from one sample to another, such as the sample average, are called *sample statistics*.

At this point, it's helpful to switch from $Avg_n[Y_i]$ to a more compact notation for averages, $\bar{Y}$. Note that we're dispensing with the subscript n to avoid clutter—henceforth, it's on you to remember that sample averages are computed in a sample of a particular size. The sample average, $\bar{Y}$, is a good estimator of $E[Y_i]$ (in statistics, an *estimator* is any function of sample data used to estimate parameters). For one thing, the LLN tells us that in large samples, the sample average is likely to be very close to the corresponding population mean. A related property is that the expectation of $\bar{Y}$ is also $E[Y_i]$. In other words, if we were to draw infinitely many random samples, the average of the resulting $\bar{Y}$ across draws would be the underlying population mean. When a sample statistic has expectation equal to the corresponding population parameter, it's said to be an *unbiased estimator* of that parameter. Here's the sample mean's unbiasedness property stated formally:

UNBIASEDNESS OF THE SAMPLE MEAN $E[\bar{Y}] = E[Y_i]$

The sample mean should not be expected to be bang on the corresponding population mean: the sample average in one sample might be too big, while in other samples it will be too small. Unbiasedness tells us that these deviations are not systematically up or down; rather, in repeated samples they average out to zero. This unbiasedness property is distinct from the LLN, which says that the sample mean gets closer and closer to the population mean as the sample size grows. Unbiasedness of the sample mean holds for samples of any size.

Measuring Variability

In addition to averages, we're interested in variability. To gauge variability, it's customary to look at average squared deviations from the mean, in which positive and negative gaps get equal weight. The resulting summary of variability is called *variance*.

The *sample variance of Y_i in a sample of size n* is defined as

$$S(Y_i)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2.$$

The corresponding *population variance* replaces averages with expectations, giving:

$$V(Y_i) = E\left[\left(Y_i - E[Y_i]\right)^2\right].$$

Like $E[Y_i]$, the quantity $V(Y_i)$ is a fixed feature of a population —a parameter. It's therefore customary to christen it in Greek: $V(Y_i) = \sigma_Y^2$, which is read as "sigma-squared-y."[16]

Because variances square the data they can be very large. Multiply a variable by 10 and its variance goes up by 100. Therefore, we often describe variability using the square root of the variance: this is called the *standard deviation,* written σ_Y. Multiply a variable by 10 and its standard deviation increases by 10. As always, the population standard deviation, σ_Y, has a sample counterpart $S(Y_i)$, the square root of $S(Y_i)^2$.

Variance is a descriptive fact about the distribution of Y_i. (Reminder: the *distribution* of a variable is the set of values the variable takes on and the relative frequency that each value is observed in the population or generated by a random process.) Some variables take on a narrow set of values (like a dummy variable indicating families with health insurance), while others (like income) tend to be spread out with some very high values mixed in with many smaller ones.

It's important to document the variability of the variables you're working with. Our goal here, however, goes beyond

[16] Sample variances tend to underestimate population variances. Sample variance is therefore sometimes defined as

$$S(Y_i)^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2,$$

that is, dividing by $n - 1$ instead of by n. This modified formula provides an unbiased estimate of the corresponding population variance.

this. We're interested in quantifying the variance of the sample mean in repeated samples. Since the expectation of the sample mean is $E[Y_i]$ (from the unbiasedness property), the population variance of the sample mean can be written as

$$V(\bar{Y}) = E\left[\left(\bar{Y} - E[\bar{Y}]\right)^2\right] = E\left[\left(\bar{Y} - E[Y_i]\right)^2\right].$$

The variance of a statistic like the sample mean is distinct from the variance used for descriptive purposes. We write $V(\bar{Y})$ for the variance of the sample mean, while $V(Y_i)$ (or σ_Y^2) denotes the variance of the underlying data. Because the quantity $V(\bar{Y})$ measures the variability of a sample statistic in repeated samples, as opposed to the dispersion of raw data, $V(\bar{Y})$ has a special name: *sampling variance*.

Sampling variance is related to descriptive variance, but, unlike descriptive variance, sampling variance is also determined by sample size. We show this by simplifying the formula for $V(\bar{Y})$. Start by substituting the formula for $\bar{Y}$ inside the notation for variance:

$$V(\bar{Y}) = V\left(\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right]\right).$$

To simplify this expression, we first note that random sampling ensures the individual observations in a sample are not systematically related to one another; in other words, they are statistically independent. This important property allows us to take advantage of the fact that the variance of a sum of statistically independent observations, each drawn randomly from the same population, is the sum of their variances. Moreover, because each Y_i is sampled from the same population, each draw has the same variance, σ_Y^2. Finally, we use the property that the variance of a constant (like $1/n$) times Y_i is the square of this constant times the variance of Y_i. From these considerations, we get

$$V(\bar{Y}) = V\left(\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right]\right) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma_Y^2.$$

Simplifying further, we have

$$V(\bar{Y}) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma_Y^2 = \frac{n\sigma_Y^2}{n^2} = \frac{\sigma_Y^2}{n}. \tag{1.5}$$

We've shown that the sampling variance of a sample average depends on the variance of the underlying observations, σ_Y^2, and the sample size, n. As you might have guessed, more data means less dispersion of sample averages in repeated samples. In fact, when the sample size is very large, there's almost no dispersion at all, because when n is large, σ_Y^2/n is small. This is the LLN at work: as n approaches infinity, the sample average approaches the population mean, and sampling variance disappears.

In practice, we often work with the standard deviation of the sample mean rather than its variance. The standard deviation of a statistic like the sample average is called its *standard error*. The standard error of the sample mean can be written as

$$SE(\bar{Y}) = \frac{\sigma_Y}{\sqrt{n}}. \tag{1.6}$$

Every estimate discussed in this book has an associated standard error. This includes sample means (for which the standard error formula appears in equation (1.6)), differences in sample means (discussed later in this appendix), regression coefficients (discussed in Chapter 2), and instrumental variables and other more sophisticated estimates. Formulas for standard errors can get complicated, but the idea remains simple. The standard error summarizes the variability in an estimate due to random sampling. Again, it's important to avoid confusing standard errors with the standard deviations of the underlying variables; the two quantities are intimately related yet measure different things.

One last step on the road to standard errors: most population quantities, including the standard deviation in the numerator of (1.6), are unknown and must be estimated. In practice,

therefore, when quantifying the sampling variance of a sample mean, we work with an *estimated standard error*. This is obtained by replacing σ_Y with $S(Y_i)$ in the formula for $SE(\bar{Y})$. Specifically, the estimated standard error of the sample mean can be written as

$$\hat{SE}(\bar{Y}) = \frac{S(Y_i)}{\sqrt{n}}.$$

We often forget the qualifier "estimated" when discussing statistics and their standard errors, but that's still what we have in mind. For example, the numbers in parentheses in Table 1.4 are estimated standard errors for the relevant differences in means.

The t-Statistic and the Central Limit Theorem

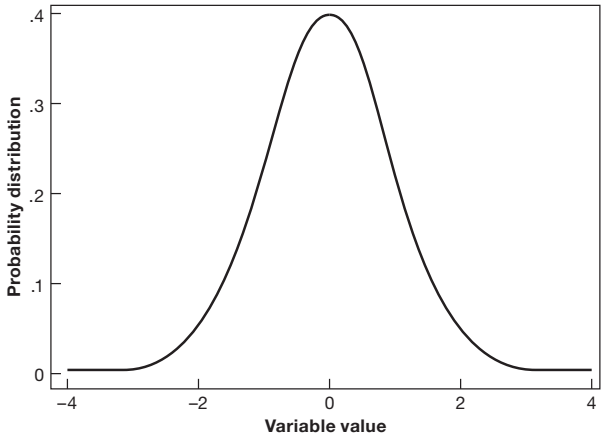
Having laid out a simple scheme to measure variability using standard errors, it remains to interpret this measure. The simplest interpretation uses a *t-statistic*. Suppose the data at hand come from a distribution for which we believe the population mean, $E[Y_i]$, takes on a particular value, μ (read this Greek letter as "mu"). This value constitutes a working *hypothesis*. A t-statistic for the sample mean under the working hypothesis that $E[Y_i] = \mu$ is constructed as

$$t(\mu) = \frac{\bar{Y} - \mu}{\hat{SE}(\bar{Y})}.$$

The working hypothesis is a reference point that is often called the *null hypothesis*. When the null hypothesis is $\mu = 0$, the t-statistic is the ratio of the sample mean to its estimated standard error.

Many people think the science of statistical inference is boring, but in fact it's nothing short of miraculous. One miraculous statistical fact is that if $E[Y_i]$ is indeed equal to μ, then—as long as the sample is large enough—the quantity $t(\mu)$ has a sampling distribution that is very close to a bell-shaped standard normal distribution, sketched in Figure 1.1. This property, which applies regardless of whether Y_i itself is normally distributed, is called the *Central Limit Theorem* (CLT). The

FIGURE 1.1
A standard normal distribution



CLT allows us to make an empirically informed decision as to whether the available data support or cast doubt on the hypothesis that $E[Y_i]$ equals μ.

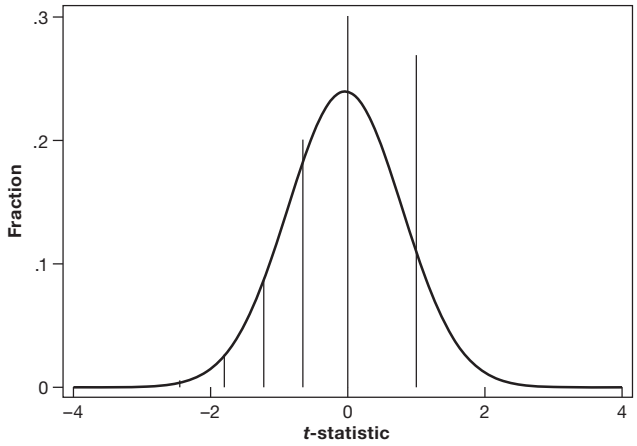
The CLT is an astonishing and powerful result. Among other things, it implies that the (large-sample) distribution of a t-statistic is independent of the distribution of the underlying data used to calculate it. For example, suppose we measure health status with a dummy variable distinguishing healthy people from sick and that 20% of the population is sick. The distribution of this dummy variable has two spikes, one of height .8 at the value 1 and one of height .2 at the value 0. The CLT tells us that with enough data, the distribution of the t-statistic is smooth and bell-shaped even though the distribution of the underlying data has only two values.

We can see the CLT in action through a sampling experiment. In sampling experiments, we use the random number generator in our computer to draw random samples of different sizes over and over again. We did this for a dummy variable that equals one 80% of the time and for samples of size 10, 40, and 100. For each sample size, we calculated the t-statistic in half a million random samples using .8 as our value of μ.

Figures 1.2–1.4 plot the distribution of 500,000 t-statistics calculated for each of the three sample sizes in our experiment, with the standard normal distribution superimposed. With only 10 observations, the sampling distribution is spiky, though the outlines of a bell-shaped curve also emerge. As the sample size increases, the fit to a normal distribution improves. With 100 observations, the standard normal is just about bang on.

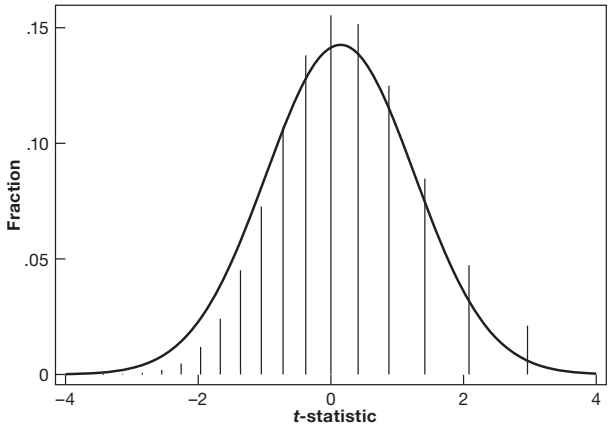
The standard normal distribution has a mean of 0 and standard deviation of 1. With any standard normal variable, values larger than ± 2 are highly unlikely. In fact, realizations larger than 2 in absolute value appear only about 5% of the time. Because the t-statistic is close to normally distributed, we similarly expect it to fall between about ± 2 most of the time. Therefore, it's customary to judge any t-statistic larger than about 2 (in absolute value) as too unlikely to be consistent with the null hypothesis used to construct it. When the null hypothesis is $\mu = 0$ and the t-statistic exceeds 2 in absolute value, we say the sample mean is *significantly different from*

FIGURE 1.2
The distribution of the t-statistic for the mean in a sample of size 10



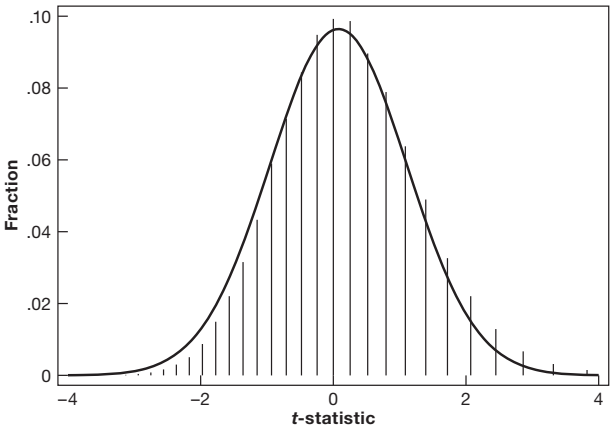
Note: This figure shows the distribution of the sample mean of a dummy variable that equals 1 with probability .8.

FIGURE 1.3
The distribution of the *t*-statistic for the mean in a sample of size 40



Note: This figure shows the distribution of the sample mean of a dummy variable that equals 1 with probability .8.

FIGURE 1.4
The distribution of the *t*-statistic for the mean in a sample of size 100



Note: This figure shows the distribution of the sample mean of a dummy variable that equals 1 with probability .8.

zero. Otherwise, it's not. Similar language is used for other values of μ as well.

We might also turn the question of statistical significance on its side: instead of checking whether the sample is consistent with a specific value of μ, we can construct the set of all values of μ that are consistent with the data. The set of such values is called a *confidence interval* for $E[Y_i]$. When calculated in repeated samples, the interval

$$\left[\bar{Y} - 2 \times \hat{SE}(\bar{Y}), \bar{Y} + 2 \times \hat{SE}(\bar{Y})\right]$$

should contain $E[Y_i]$ about 95% of the time. This interval is therefore said to be a *95% confidence interval* for the population mean. By describing the set of parameter values consistent with our data, confidence intervals provide a compact summary of the information these data contain about the population from which they were sampled.

Pairing Off

One sample average is the loneliest number that you'll ever do. Luckily, we're usually concerned with two. We're especially keen to compare averages for subjects in experimental treatment and control groups. We reference these averages with a compact notation, writing $\bar{Y}^1$ for $Avg_n[Y_i|D_i = 1]$ and $\bar{Y}^0$ for $Avg_n[Y_i|D_i = 0]$. The treatment group mean, $\bar{Y}^1$, is the average for the n_1 observations belonging to the treatment group, with $\bar{Y}^0$ defined similarly. The total sample size is $n = n_0 + n_1$.

For our purposes, the difference between $\bar{Y}^1$ and $\bar{Y}^0$ is either an estimate of the causal effect of treatment (if Y_i is an outcome), or a check on balance (if Y_i is a covariate). To keep the discussion focused, we'll assume the former. The most important null hypothesis in this context is that treatment has no effect, in which case the two samples used to construct treatment and control averages come from the same population. On the other hand, if treatment changes outcomes, the populations from which treatment and control observations are

drawn are necessarily different. In particular, they have different means, which we denote μ^1 and μ^0.

We decide whether the evidence favors the hypothesis that $\mu^1 = \mu^0$ by looking for statistically significant differences in the corresponding sample averages. Statistically significant results provide strong evidence of a treatment effect, while results that fall short of statistical significance are consistent with the notion that the observed difference in treatment and control means is a chance finding. The expression "chance finding" in this context means that in a hypothetical experiment involving very large samples—so large that any sampling variance is effectively eliminated—we'd find treatment and control means to be the same.

Statistical significance is determined by the appropriate t-statistic. A key ingredient in any t recipe is the standard error that lives downstairs in the t ratio. The standard error for a comparison of means is the square root of the sampling variance of $\bar{Y}^1 - \bar{Y}^0$. Using the fact that the variance of a difference between two statistically independent variables is the sum of their variances, we have

$$\begin{aligned} V\left(\bar{Y}^1 - \bar{Y}^0\right) &= V\left(\bar{Y}^1\right) + V\left(\bar{Y}^0\right) \\ &= \frac{\sigma_Y^2}{n_1} + \frac{\sigma_Y^2}{n_0} = \sigma_Y^2 \left[\frac{1}{n_1} + \frac{1}{n_0}\right]. \end{aligned}$$

The second equality here uses equation (1.5), which gives the sampling variance of a single average. The standard error we need is therefore

$$SE\left(\bar{Y}^1 - \bar{Y}^0\right) = \sigma_Y \sqrt{\frac{1}{n_1} + \frac{1}{n_0}}.$$

In deriving this expression, we've assumed that the variances of individual observations are the same in treatment and control groups. This assumption allows us to use one symbol, σ_Y^2, for the common variance. A slightly more complicated formula allows variances to differ across groups even if the means are

the same (an idea taken up again in the discussion of robust regression standard errors in the appendix to Chapter 2).[17]

Recognizing that σ_Y^2 must be estimated, in practice we work with the estimated standard error

$$\hat{SE}\left(\bar{Y}^1 - \bar{Y}^0\right) = S(Y_i)\sqrt{\frac{1}{n_1} + \frac{1}{n_0}}, \qquad (1.7)$$

where $S(Y_i)$ is the *pooled sample standard deviation*. This is the sample standard deviation calculated using data from both treatment and control groups combined.

Under the null hypothesis that $\mu^1 - \mu^0$ is equal to the value μ, the t-statistic for a difference in means is

$$t(\mu) = \frac{\bar{Y}^1 - \bar{Y}^0 - \mu}{\hat{SE}\left(\bar{Y}^1 - \bar{Y}^0\right)}.$$

We use this t-statistic to test working hypotheses about $\mu_1 - \mu_0$ and to construct confidence intervals for this difference. When the null hypothesis is one of equal means ($\mu = 0$), the statistic $t(\mu)$ equals the difference in sample means divided by the estimated standard error of this difference. When the t-statistic is large enough to reject a difference of zero, we say the estimated difference is statistically significant. The confidence interval for a difference in means is the difference in sample means plus or minus two standard errors.

Bear in mind that t-statistics and confidence intervals have little to say about whether findings are substantively large or small. A large t-statistic arises when the estimated effect of interest is large but also when the associated standard error is small (as happens when you're blessed with a large sample). Likewise, the width of a confidence interval is determined by

[17] Using separate variances for treatment and control observations, we have

$$SE\left(\bar{Y}^1 - \bar{Y}^0\right) = \sqrt{\frac{V^1(Y_i)}{n_1} + \frac{V^0(Y_i)}{n_0}},$$

where $V^1(Y_i)$ is the variance of treated observations, and $V^0(Y_i)$ is the variance of control observations.

statistical precision as reflected in standard errors and not by the magnitude of the relationships you're trying to uncover. Conversely, *t*-statistics may be small either because the difference in the estimated averages is small or because the standard error of this difference is large. The fact that an estimated difference is not significantly different from zero need not imply that the relationship under investigation is small or unimportant. Lack of statistical significance often reflects lack of statistical precision, that is, high sampling variance. Masters are mindful of this fact when discussing econometric results.