CHAPTER 1

Introduction

The writings of Thomas Malthus and David Ricardo earned economics the nickname the dismal science. In *An Essay on the Principle of Population* (1798), Malthus argued that population growth inevitably outstrips the growth of food production, so eventually population becomes too great for food supply. In *The Principles of Political Economy and Taxation* (1817), Ricardo expanded this argument to the "iron law of wages." At wages above the subsistence level, population grows, driving down wages. At wages below the subsistence level, the poor face starvation, population declines, and wages rise. Poverty policies that transfer money to the poor must be ineffective in the long run because wages must always end up at the subsistence level.

Writing less than a century later, Alfred Marshall was much more optimistic. He believed it was possible to eliminate poverty, if not within a generation, at least within two generations.[1] Marshall thought poverty policy consisted of increasing the demand for labor and reducing the supply of unskilled labor. He argued that economic growth increased the demand for labor. Increased education would lower the supply of unskilled labor, both directly, by moving workers from the unskilled to the skilled labor force, and indirectly, by reducing population growth. To further reduce the size of the unskilled labor force, society should encourage later marriages and childbearing among the "lower strains." Finally, government should address the environment in which the poor lived. Marshall commented on the benefits of the suburbs, "where excellent systems of drainage, water supply and lighting, together with good schools and opportunities for open air play, give conditions at least as conducive to vigour as are to be found in the country," and continued, "There is no better use for public and private money than in providing public parks and playgrounds in large cities, in contracting with railways to increase the number of workmen's trains run by them, and in helping

1. Alfred Marshall, "Three Lectures on Progress and Prosperity," 1883; reprinted in *Journal of Law and Economics* 12 (April 1969): 184–212.

those of the working classes who are willing to leave the large towns to do so, and to take their industries with them."[2] Marshall saw little role for redistribution of income to the poor. Giving the poor money would reduce their industriousness. Only the victims of misfortune were good candidates for such charity.

In sharp contrast, Henry George, in *Progress and Poverty* (1879), maintained that because land was fixed in supply and necessary for production, only landowners would benefit from progress. According to George, the solution was to raise taxes on land (not on structures or other improvements) in proportion to their value. In this way, the benefits from the increased value of land could be shared with workers.

1. The Content of This Book

From the vantage point of another century of experience, it is clear that Marshall's optimism was in part justified and in part exaggerated. Standards of living are much higher than they were in the late nineteenth century. On the other hand, poverty did not disappear in two generations. In large part, this is because our understanding of what it means to be poor has changed over time. Much of chapter 2 is devoted to exploring this issue.

But the issue is not only that we have redefined who is poor. Using a constant definition of poverty, over the past three decades there has been little change in the proportion of Americans who are poor, despite dramatic increases in average incomes. Thus the Marshall-George debate remains relevant. Has the relation between economic growth and poverty broken down? Did some other trend hide the positive effects of economic growth? Or is the problem the way we measure poverty? Chapters 3 and 4 address these issues.

Many of Marshall's concerns remain relevant today. As chapter 4 shows, there is a strong relation between poverty and the state of the labor market for low-wage workers. Chapter 5 addresses the effectiveness of different policies designed to raise after-tax wages for low-wage workers. Some of these policies follow in the Marshallian tradition of increasing skills and demand for low-wage workers. Others have a somewhat more Georgist flavor (although certainly not based on taxing land), using the tax system to support low-wage workers or intervening directly in the wage-setting process.

We also see renewed focus on the family and on increasing the age of mothers at first childbirth. Chapter 6 addresses issues such as out-of-wedlock births and teenage childbearing. It looks at policies designed to help children. As discussed earlier, Marshall also believed that poverty was closely related to place. He advocated getting the poor out of crowded cities. Today, many analysts believe that concentrated poverty is a particular problem. Chapter 7 addresses the issues associated with such poverty and programs designed to alleviate it.

Since Marshall's time, the availability of public schooling has increased dramatically. Not only public primary schooling but public secondary schooling is universally

2. Marshall, 199–200.

available in the United States. Yet the effectiveness of public education, particularly in high-poverty areas, is hotly debated in this country. Chapter 8 discusses the debate over education reform.

Finally, the concern that programs that support the poor will "sap their industriousness" has been a recurring theme over the centuries of poverty policy. The 1996 Personal Responsibility and Work Opportunity Reconciliation Act, more commonly known as welfare reform, was a response to concerns that welfare was hurting the very people it was designed to help, very much as Marshall believed that the Poor Laws passed in the late eighteenth century in England had hurt laborers by encouraging them to rely on support for the poor. It is still somewhat early to assess the impact of welfare reform, but chapter 9 discusses the background of reform and what we know about its effects.

Although most of this book is about poverty, the last third is about discrimination, and of that last third, three of the four chapters focus on race discrimination while the last is concerned with sex discrimination. There are two reasons for combining discussions of poverty and discrimination in a single book. The first is that the methods researchers use to study the two topics are closely related. Discrimination research relies on a combination of theory, observational studies, and experimental methods similar to those used in poverty research (discussed later). Although discrimination research relies more heavily on theory and less heavily on actual or quasi-experiments than does poverty research, it is easy to make the transition from the latter to the former. For the most part, those of you reading this book do not need a new set of analytical tools. Indeed, when we discuss sex discrimination, we will return to some of the same theories that we will have discussed when analyzing the relation between the decline of marriage and poverty.

The second reason for discussing poverty and discrimination in the same book is that, although the topics are distinct, they are also related. Most of the poor are not black, but the poverty rate is much higher among black Americans than among white Americans. To the extent that discrimination contributes to lower incomes among blacks, it contributes to poverty and helps to account for their higher poverty rate.

But it is also likely that higher poverty rates among blacks contribute to discrimination. Chapter 10 discusses a variety of theories of discrimination. Most rest on perceived or actual differences (or both) between blacks and whites. If blacks tend to come from more disadvantaged backgrounds than do whites, this can affect the flow of information from potential employees to potential employers and can, through multiple mechanisms, reduce the employment prospects of blacks relative to whites, even whites from the same background.

On the other hand, these differences could promote prejudice, but in many settings people are unable to act on their prejudice. Therefore, the existence of prejudice need not lead to worse outcomes for blacks than for otherwise equivalent whites. Chapter 11 examines the evidence for and against the existence of discrimination in the labor market as well as the role of policy in addressing labor market discrimination. Many researchers believe that differences between blacks and whites in labor market outcomes primarily reflect differences in the skills that people bring to the labor

market. Chapter 12 explores the black-white test score gap and issues regarding differences in access to schooling and desegregation. Chapter 13 reviews the evidence on discrimination in other domains, including the justice system and customer markets.

Just as poverty is more common among blacks than among whites, it is more common among women than among men and is particularly common among female-headed households. Thus differences in earnings capacity between men and women may have an important impact on the prevalence of poverty. Chapter 14 addresses the debate over the source of this differential and policies designed to reduce it.

Some readers of earlier drafts of this book have commented that it would benefit from a lengthier discussion of inequality. After all, one important explanation for the lack of a decline in poverty over the past three decades is the dramatic increase in inequality. In the concluding chapter, I argue that reducing inequality must be an important component of any policy that reduces poverty. Nevertheless, I have made only a modest attempt to accommodate readers who have requested more on this subject.

There are two reasons that I have not added a complete chapter or more on the study of inequality. The first is simply a matter of space and time. This book is already longer than either the publisher or I anticipated, and it may well contain more material than can be covered in a typical semester course (although I do cover most of the material in a semester).

The second reason is that the study of inequality in many ways relies on a different set of tools and methods than does the study of poverty and discrimination. By its nature, the theory of inequality requires a more global approach. The empirical analysis of inequality, for the most part, uses a different statistical approach and, in particular, makes less use of actual and quasi-experiments, which are the focus of much of this book.

2. Recent Developments in the Study of Poverty and Discrimination

The earlier discussion may have given the impression that little has changed since Marshall and George debated progress and poverty. In fact, the study of poverty in general and poverty policy in particular has changed dramatically over the past thirty years. A large part of the impetus for the change can be traced to the debate over the negative income tax.

In 1962, the future Nobel laureate Milton Friedman proposed that the welfare system in the United States be replaced with a negative income tax.[3] Under a negative income tax, all individuals or households with incomes below a certain level would receive a basic guaranteed annual income from the government. As household income increased, the government grant (or negative income tax) would be scaled back. When household income was sufficiently high, the household would not receive any grant from the government but instead would pay income tax as it would in a standard income tax system. Although Friedman was a well-known conservative, he was also a highly

3. Milton Friedman, *Capitalism and Freedom* (Chicago: Chicago University Press, 1962).

respected economist, and his proposal was subject to considerable analysis in professional journals.

Support for the negative income tax crossed political boundaries. The first proposal for a negative income tax came from the Johnson administration in 1965, and the 1967 reforms to the welfare system reflect some of the spirit of the negative income tax proposal. Nixon's Family Assistance Plan proposed a form of negative income tax that was opposed by welfare rights activists on the left.[4] James Tobin, also a Nobel laureate but, in contrast with Friedman, a recognized liberal, designed a negative income tax proposal for the McGovern campaign in 1972.[5]

The key aspects of the theoretical analysis can be summarized briefly. If the negative income tax were to be affordable, the rate at which the grant would be reduced as income increased would have to be substantial. At a minimum, the grant would decline by one dollar for every three dollars of income and more probably by one dollar for every two dollars of income. In today's terms, even using the lower rate, a worker earning nine dollars an hour would see his family's grant fall by three dollars for each hour that he worked. Thus, in effect, he would be earning only six dollars an hour. Because the after-tax hourly wage of working families would be much lower but their overall income would be higher, these families might work less under a negative income tax than they would otherwise.[6]

On the other hand, under traditional welfare, benefits were typically reduced by one dollar for every one dollar a recipient earned. Thus people with very low potential earnings who were therefore unlikely to earn much more than they would receive from welfare had little or no incentive to work. Under a negative income tax, they would keep some of their earnings and thus have some incentive to work. The advocates of the negative income tax hoped that it would encourage very low-income families to work and not reduce labor supply very much among somewhat higher-income households.

But of course it was possible that just the opposite would happen. Perhaps current welfare recipients would be little affected if their incentive to work were increased and there would be a big reduction in work effort among near-poor families who did not receive assistance through traditional welfare. This issue could not be resolved on the basis of theory alone.

Economists began with observational studies, that is, they looked at the relation between after-tax wages and labor supply in the population. They relied on surveys of individuals who reported, among other variables, their hours of work and earnings or wage rate.

We can think of an observational study in the following way. Suppose we find a sample of people who seem to be otherwise similar but some of whom earn six dollars

4. Walter Williams, "The Continuing Struggle for a Negative Income Tax: A Review Article," *Journal of Human Resources* 10 (Fall 1975): 427–44.

5. Holcomb B. Noble, "James Tobin, Nobel Laureate in Economics and an Adviser to Kennedy, Dies at 84," *New York Times,* March 13, 2002.

6. For a formal discussion of the effects of income and wage rates on labor supply, see the discussion of the earned income tax credit in chapter 3.

an hour and some of whom earn nine dollars an hour. If the only difference between the two groups is that some people were lucky and got jobs paying nine dollars an hour and others were unlucky and got jobs paying six dollars hour, it may be reasonable to assume that if we cut the pay of the lucky people to six dollars an hour, they would act like the unlucky people. Suppose that similar people earning six dollars an hour worked two hours per week less than those earning nine dollars an hour. Then we might conclude that if we were to tax workers earning nine dollars an hour so that they ended up earning six dollars an hour, they would reduce their labor by two hours per week.

The implicit assumption that the only difference between the groups is how lucky they were is very strong. It is likely that even though they look similar on paper, the people earning nine dollars an hour are somehow different from those earning six dollars an hour. For example, they may be more skilled or work harder even though they have similar educations. Or their jobs may be different. The higher-paying job may be more dangerous or more demanding. So the real challenge for the statistician, and one to which we will devote a great deal of time in this book, is figuring out ways to obtain samples of people who differ only along the dimension we are trying to study.

A classic book by Glen Cain and Harold Watts brought together seven papers focused on predicting the effect of a negative income tax on labor supply.[7] In their conclusion, Cain and Watts point out the large range in the estimated effects of very similar programs. One study found that a $3,000 guarantee coupled with a 50 percent tax rate would have a negligible effect on the labor supply of husbands. In contrast, a second study predicted that a less generous program with a $2,400 guarantee and a 50 percent tax rate would reduce the labor supply of male family members by 37 percent. The former study implied that the negative income tax would be a cost-effective approach to reducing poverty. The latter implied that it would be very costly.

As reflected in this example, it is often very difficult to use observational data to obtain convincing evidence of the causal effect of one variable (such as the after-tax wage or parental absence) on a second variable (such as labor supply or adult outcomes). If we could conduct an experiment in which we randomly assigned some people to have high wages and some people to have low wages, we would have much more convincing evidence regarding the relation between labor supply and after-tax wages.

Because the stakes involved in instituting a negative income tax were so high, policy analysts convinced the federal government to conduct experiments in which some people were randomly assigned to be eligible for the negative income tax while others were randomly assigned to remain subject to traditional welfare. In the experimental group, there was also variation in the generosity of the program. Four experiments were conducted in the United States and a fifth in Canada. The largest of these is known as SIME/DIME (the Seattle Income Maintenance Experiment / Denver Income Maintenance Experiment).

Many policy analysts found the results of the experiments disappointing. Although the labor supply response was modest, it added substantially to the cost of the program.

7. Glen G. Cain and Harold W. Watts, *Income Maintenance and Labor Supply* (New York: Academic Press, 1973).

Depending on the generosity of the program evaluated in SIME/DIME, the labor supply response could account for over half of the costs.[8] The least generous program would save $4 billion but would make 95 percent of recipients worse off. A program that would guarantee support at the poverty level and tax-back benefits at a rate of only 50 percent would still make one-fourth of recipients worse off and would exceed the cost of the welfare program then current by $30 billion, an enormous increase.[9]

There were also some "unintended consequences," the social science equivalent of medical side effects. In particular, the divorce rate rose among recipients randomly assigned to the negative income tax.[10] The combination of the costs of the labor supply effects and the effect on marriage led Senator Daniel Patrick Moynihan, an early supporter of the negative income tax, to withdraw his support.

Despite the failure of the policy proposal, the negative income tax experiments established the value of large-scale experiments and helped to create the infrastructure to carry them out. Randomized experiments have become the "gold standard" by which empirical work on poverty and discrimination is judged. The welfare reform passed by Congress in 1996 was heavily influenced by the outcomes of a large number of experiments that evaluated potential reforms.

Of course, it is not possible to use experiments to evaluate all policy proposals. Some experiments would be immoral if not impossible to perform. And a badly conducted experiment is still bad research even though that research uses an experiment. Other experiments are simply too expensive compared with the benefit they are expected to bring. In such cases, researchers can sometimes rely on what are called "natural experiments" or "quasi-experiments."

These experiments attempt to find situations that mimic randomized trials. In these situations, it is as if the researcher had randomly assigned participants to a treatment or control group. Participants are said to be as good as randomly assigned. We will discuss later many of the issues associated with particular quasi-experiments. In general, however, we will be concerned with whether participants are truly "as good as randomly assigned" and with whether the control group may be affected by the experiment. The answers to these questions are frequently unclear even in the case of true experiments but especially when we examine quasi-experiments.

Despite the difficulties associated with particular experimental or quasi-experimental studies, such studies, when done well, are generally more convincing than observational studies. This does not mean that there is no longer a role for observational studies in

8. Michael C. Keeley, Philip K. Robins, Robert G. Spiegelman, and Richard W. West, "The Labour Supply Effects and Costs of Alternative Negative Income Tax Programmes," *Journal of Human Resources* 13 (Winter 1978): 3–36.

9. Robert G. Spiegelman and K. E. Yaeger, "The Seattle and Denver Income Maintenance Experiments: Overview," *Journal of Human Resources* 15 (Fall 1980): 463–79.

10. Lyle P. Groeneveld, Nancy B. Tuma, and Michael T. Hannan, "The Effects of Negative Income Tax Programs on Marital Dissolution," *Journal of Human Resources* 15 (Fall 1980): 654–74; but see Glen G. Cain and Douglas A. Wissoker, "A Reanalysis of Marital Stability in the Seattle-Denver Income-Maintenance Experiment," *American Journal of Sociology* 95 (March 1990): 1235–69, for a reanalysis challenging this conclusion.

research on poverty and discrimination. Observational studies are often the basis for justifying more careful experimental study. Or, if it is impossible to design a convincing experiment or quasi-experiment, we may be forced to fall back on an observational study.

Moreover, for several reasons, experiments are only imperfect substitutes for theory. First, theory is an important guide to which experiments to conduct and how to interpret them. Second, suppose we learn (as we do in chapter 11) that if they apply randomly for jobs, individuals with names that reveal that they are black are less likely to be called for an interview. Theory will tell us that the effect of this discriminatory behavior depends a great deal on how the labor market works. If applicants have a good idea of which firms discriminate and which do not, they will not apply randomly, and the experiment will not give a clear impression of the effect of this discriminatory behavior on applicants. We might follow up the initial experiment with a study of how black applicants decide where to apply for jobs and, in particular, see whether they avoid the firms for which we have evidence of discriminatory behavior.

By now it may occur to you that the study of poverty and discrimination increasingly resembles the study of medicine in its reliance on experiments. It would be false to view medicine as atheoretical. Theory guides the choice of medicines and medical procedures to study. Nor is empirical analysis in medicine purely experimental. We learn a great deal from observational studies. Often observational studies justify experimental studies. But ultimately, experimental studies are the strongest guide to good practice. New medicines are subject to clinical trials in which outcomes from those receiving the medicine are compared with outcomes from those receiving either no treatment or the standard treatment. And researchers watch for unintended consequences in the form of side effects.

Research on poverty and discrimination certainly does not rival experimental medical research, let alone research in microbiology. This reflects in part the difficulty of working with human beings who can see through experiments. A participant in a negative income tax experiment who knows that the experiment will last only three years need not respond to the experiment in the same way that she would if a negative income tax were established permanently. In contrast, we expect that the body's response to medication is not affected by the participant's understanding that the experiment is of limited duration (although willingness to take the medication consistently may be affected).

Research in both medicine and the social sciences suffers from the fact that the environment changes in response to our policies. Getting a flu shot reduces not only the chance of getting the flu but also the probability of spreading it to others. Thus the effect of a policy making flu shots freely available would be different from the effect that would be predicted by a small-scale experiment. Similarly, with a universal negative income tax, the decline in labor force participation would affect the availability of jobs and the wages they paid. Again, this means that we cannot dispense with theory.

3. The Object of This Book

What do we know about poverty and discrimination? And how do we know it? The body of literature on this topic is enormous. Unfortunately, much of what has been

written is designed to further a political agenda, and much of the rest is just not very good.

The goal of this book is to help you distinguish the good research from the rest. I discuss in detail a small number of the best studies on each topic. This approach will allow you to understand not only the principal findings of the study but also the weaknesses that limit our confidence in its results. As the introduction to this chapter should make clear, even the best studies are imperfect. I try to avoid summarizing a large number of studies, many of which suffer from significant shortcomings. When choosing among good studies, I have tried to select the one that is most accessible.

Inevitably, on occasion I am forced to say "most researchers believe that" or "most studies show that" or "the evidence is mixed" because it would be too time-consuming to discuss the individual studies. In general, however, I try to resist the temptation to be the ultimate judge of a body of literature rather than giving you the tools and information you need to evaluate it.

You may come away from the book feeling less sure about what you know than you did before you read it. That is good, not bad. Although many people offer simple solutions to the problems of poverty and discrimination, these problems are genuinely complex. If there were simple solutions, we probably would have done away with both of these problems. After all, most people think that poverty and discrimination are bad and would like to get rid of them if they could.

We will discuss various policies that have been tried in the attempt to end poverty and/or discrimination. We will discuss whether they worked. In some cases, we will consider arguments that the very policies intended to reduce poverty and discrimination have increased them. This should give you an idea of just how hard it will be to solve these problems. There are probably no "right" answers to the questions that we will raise. But there are better and worse answers and even wrong answers.

Some people will find this conclusion depressing. A number of students have made comments to me like "How can you stand working in this area? The answer is always 'We don't know.'" But it is precisely because there is so much that we do not know and understand about such an important topic that it is fun to study.

And of course, except in some deep metaphysical sense, it is not true that we do not know anything. We know a great deal, and we have learned much of it in the past ten years. Still, it is also important to recognize what we do not know and that the evidence for what we think we know is often weak.

This book concentrates on statistical analyses and formal theories. Certainly our understanding of poverty and discrimination is informed by careful and thoughtful descriptions of the poverty experience. Reading ethnographic studies[11] gives enormous insight into the lives of poor people, and this book will draw on these insights. But we will not talk a lot about what it is like to be poor and how it feels to be the victim of discrimination. Instead, we will spend a little time doing economic theory and a lot of time looking at arguments based on statistical analysis.

11. Ethnographic studies provide rich descriptions of the functioning of human societies. Several excellent ethnographies are included in the section on additional readings at the end of this chapter.

In a sense, this book is about statistical analysis. By this I do not mean that it covers statistical techniques, although the appendix to this chapter covers the basics you need to read this book. I mean that it is about how to assess the quality of statistical arguments. Statistical arguments are generally imperfect. There are usually other explanations for a result than the one presented by a researcher. Sometimes those other explanations are equally or even more plausible than the one presented.

Understanding how to assess statistical arguments is a skill that is valuable well beyond the study of poverty and discrimination. Businessmen, doctors, policy analysts, and many others rely on statistical analysis. Understanding the quality of this analysis may be important for making the right business decision or choosing the right medical approach. Understanding the limits of statistical analysis can also help you understand why policy analysts disagree about policies. I refer to "policy analysts." I could say "social scientists" or "economists," but the poverty field is highly policy oriented, and this book follows in that tradition.

4. Why Do Policy Analysts Disagree? The Limits of Statistical Arguments

Ultimately statistical analysis is about correlation, the degree to which characteristics tend to vary together: more educated people tend to have higher incomes than do less educated people; teenage mothers tend to have lower incomes than women who did not have children as teenagers; users of the leading asthma medicine are more likely to have bad asthma than people who do not use that medicine.

It is the policy analyst who interprets these correlations as showing a causal relation: getting more education increases income; having a child as a teenager lowers a woman's income; using the leading asthma medicine worsens rather than improves asthma. But the opposite interpretations are also possible: people who are going to earn more money do not feel as much pressure to start working soon and thus get more education; women who expect to have low earnings are more likely to become teenage mothers; people who have bad asthma are more likely to use the leading asthma medicine. Or some other factor may cause both characteristics: smart people tend to choose to get more education, and smart people tend to have higher earnings; women from disadvantaged backgrounds are more likely to become single mothers and are more likely to have low incomes; doctors who treat people who live in the Bronx like to prescribe the leading asthma medicine, and people who live in the Bronx are more likely than other Americans to have asthma.

In each case, it may seem obvious to you which of these explanations is correct. I certainly believe that the most likely reason that heavy users of the leading asthma medicine are more likely to have asthma is that asthmatics are the people most likely to use the medicine. But I cannot prove that my belief is correct simply by pointing to the relation between medicine use and asthma. If another policy analyst is convinced that use of the medicine and asthma are related because the medicine causes (or worsens) asthma, we will have to look for new evidence to help us distinguish between the two explanations.

Perhaps we can find an experiment in which a researcher randomly gave some people the asthma medicine and others a placebo. If we find that asthma was no more common among those receiving the medicine than among those who received the placebo, we may feel justified in concluding that the medicine does not cause asthma and that my original belief was correct: people who use the leading asthma medicine are more likely to have asthma than are people who do not because people with asthma are more likely to use the medicine. But the experiment is not definitive. Perhaps the reason that asthma is no more common among those who get the medicine than among those who get the placebo is because the placebo also causes asthma.

Policy analysts disagree in part because they have different beliefs about how likely these alternative explanations are. All policy analysts work with implicit models of how the world works that reflect both their life experience and their academic training. Social workers and economists do not necessarily view the world in the same way. Faced with the finding that teenage mothers have lower incomes than women who did not give birth as teens, one researcher may see evidence that teenage motherhood causes poverty while another sees evidence that poverty causes teenage motherhood. An economist is likely to respond that if having a child as a teenager is extremely costly, only those with strong reasons to have children as teenagers will do so. They may therefore be inclined to believe an explanation indicating that the poor economic prospects of some women lead them to have children at an early age. Child psychologists and social workers are less likely to believe that teenagers make rational decisions, which may make them more likely to believe that having a child at an early age has major negative consequences for the mother. Of course, both or neither may be true.

It would be easy for you to translate the last paragraph as "It's all a matter of opinion." I hope that you will not. There *are* weaker and stronger statistical arguments. The best way to find out which, if either, of these arguments is correct would be to conduct two experiments. In one experiment, we would randomly assign girls to poor and wealthy families and see if there was a difference in the proportions becoming teenage mothers. In the other experiment, we would force some teenage girls to become mothers and ensure that others did not. If the teenage mothers ended up with lower incomes than the other women, we would be reasonably confident that teenage motherhood lowered their incomes. If there were no difference in the earnings of the two groups, we would conclude that teenage motherhood did not cause lower incomes.

But even here there would be a problem. Perhaps some girls receive lower incomes as a result of becoming teenage mothers and others receive higher incomes. Those who will receive higher incomes become teenage mothers while those who would be hurt financially do not become teenage mothers. So our experiment would give the wrong answer, because it asks what is the effect of teenage motherhood on the average teenage girl rather than on the type of teenager who becomes a teenage mother.

Of course, both of these experiments would be totally immoral and would not be conducted by any ethical person. However, we will see that clever researchers often try to imitate experiments by comparing groups that are similar except for the factor they are examining. Perhaps we could compare twin sisters, one of whom gave birth while a teenager and the other of whom did not. A study discussed later in this book

compares women who were teenage mothers with women who had miscarriages as teenagers.

5. Why Do Policy Analysts Disagree? The Role of Values

To some extent, policy analysts disagree about policies because they disagree about their effects. We have seen that it is essentially impossible to prove the case for or against a policy based on data alone. However, it is frequently the case that policy analysts do agree about the effects.

Later in the book, we will examine the effects of minimum wage laws. There is considerable consensus regarding these effects. Most (although certainly not all) economists agree that minimum wage laws (at the levels found in the United States) reduce employment but that the effect is small. They also agree that minimum wage laws reduce wage inequality but do not have a large effect on income inequality or on the poverty rate.

Given this consensus, why do economists disagree about whether the minimum wage should be raised? One explanation is values.[12] Minimum wage laws are inefficient—they reduce employment—and most of us agree that is bad. But minimum wage laws reduce wage inequality, may reduce family income inequality, and make those people working in low-wage jobs more capable of supporting themselves. Many people, perhaps most, think that those are good things. We must now decide what weight to put on the good and bad effects. Reasonable people can arrive at different judgments.

Most policy analysts have concluded that the negative income tax is too expensive because it reduces work effort too much. However, others disagree. The philosopher and economist Philippe van Parijs argues that the reduction in work effort is a benefit, not a cost, of the negative income tax. In his view, society's objective should be to have people work only for the innate pleasure they derive from working and not for the income they receive. The fact that people work less when they can receive a subsidy from the government makes the financial cost of the negative income tax higher but also lessens the rat race.[13]

Similarly, when welfare reform was passed during the Clinton administration, most people believed that it would encourage families to leave welfare and find jobs (which most people felt was good). Most people also believed that it would leave those people who did not find jobs worse off. Part of the disagreement over the reform had to do

12. Victor R. Fuchs, Alan B. Krueger, and James M. Poterba, in "Economists' Views about Parameters, Values, and Policies: Survey Results in Labor and Public Economics," *Journal of Economic Literature* 36 (September 1998): 1387–425, examine the relation among values, beliefs about the effects of policies, and support for policies. They find that among economists specializing in labor economics and public economics, the two areas they study, policy disagreements are influenced much more by differences in values than by differences in beliefs about the effects of policies.

13. See, for example, Robert Van der Veen and Philippe van Parijs, "A Capitalist Road to Communism," *Theory and Society* 15 (September 1986): 635–55.

with how large each of those effects would be. Part reflected different values—how much some valued increasing the number of children growing up in families that were independent of welfare and how much others valued making it less likely that children would go hungry.

Put differently, almost all policy analysts would agree that if there is a way to make some people better off without making anyone worse off, we should do it. This is known as the *Pareto principle.* However, most policies do not fit nicely into this category. Policies that do not help anyone will generally not receive much support, while those that help at least some people and do not hurt anyone will be quickly enacted. What is left is a set of policies that help some people and hurt others. The policy analyst can help determine who is helped and who is hurt, and by how much and in what ways. Ultimately, however, the policy advocate (who may be the same person) takes over and makes the case that the benefits to some outweigh the costs to others or vice versa.

6. A Case Study: Retention in Grade

Writing in 1989, one analyst concluded, "There is probably no widespread educational practice as thoroughly discredited as retention in grade. If the research undercutting this practice is sound, the task is to uproot outdated misconceptions appealing to educators' 'common sense' wisdom."[14]

How did researchers "know" retention was bad, and why would policy makers not listen?

6.1. The Early Research. What we knew was that students who were retained in grade did not catch up with their peers. If anything, they fell further behind. They had lower self-esteem than other students and were more likely to drop out.

If retaining students in grade *caused* these differences, for most of us, this would make a compelling case that retention is a bad policy. However, there are good reasons for questioning whether the relation is causal. Students who struggle in school are more likely to be retained in grade. Students who have difficulty with school one year are more likely than other students to have difficulty in other years. We would not be surprised to find that these students also are more likely to drop out and to have low self-esteem even if they are not retained in grade.

To use an analogy, compared with those who receive high grades, students who receive low grades in high school do worse, on average, on the SAT (Scholastic Aptitude Test) or the ACT (American College Test). Few people would suggest that giving everyone As would improve SAT scores.

Serious researchers understood this problem. They tried to find students who were not retained in grade but who looked on paper a lot like the students who were retained in grade. They could match students on factors such as their race, sex, month

14. Roy P. Doyle, "The Resistance of Conventional Wisdom to Research Evidence: The Case of Retention in Grade," *Phi Delta Kappan* 71 (November 1989): 215–20.

of birth, region, parents' marital status, and education. They still found that being retained in grade was associated with bad outcomes.

How convincing is this? Suppose we found a sample of same-sex twins in which one twin was retained in grade and the other was not. We would still be worried that the twin who was retained in grade was more academically challenged than the one who was not and would therefore have done worse anyway. Back to our analogy. If the twin who got lower grades in school did worse on the SAT, would we blame the low grades for the lower performance on the SAT?

6.2. Recent Research. We can do better if we find a setting in which there is a relatively sharp cutoff establishing who is promoted and who is retained. Suppose a large school system sets a rule that students who score 65 or higher on a citywide test will be promoted, while all those who score 64 or less will be retained in grade. We will assume that the test is administered in a standard manner across schools and is fairly graded.

All tests have a random component. Was the student lucky or unlucky in his choice of topics to study? Did she accidentally mark the wrong box on the answer sheet? Did he make a lot of lucky guesses? For this reason, students who score 65 on the test should be a lot like students who score 64 on the test.

We can compare the future performance of students who scored 64 on the test and were therefore retained with the future performance of those who scored 65 and were therefore promoted. If retention helps students, we would expect the future scores of those with a test score of 64 to be higher. If it hurts, we would expect the opposite. If there is no difference, retention neither helps not hurts, but given its cost, it is probably not a good idea.

Of course, in the real world, the line between being promoted and being retained in grade is likely to be a little fuzzy. The district may have a waiver policy that allows some students who fail the test to be promoted anyway, and it probably has other requirements that can cause a student who passes the test to be retained. But the basic idea remains the same. If, for example, 95 percent of students who pass the test are promoted and only 20 percent of those who fail it are promoted, if retention is good, we should see better future outcomes for students with 64s than for students with 65s.[15]

Jenny Nagaoka and Melissa Roderick used essentially this approach. In Chicago, third graders who, after a summer remedial program, were more than one year below grade level on the Iowa Test of Basic Skills reading test were retained in grade. For sixth graders the cutoff was one and a half years below grade level. Nagaoka and Roderick compared students up to three-tenths of a year below this cutoff with students at this cutoff or an equivalent amount above. For both grades, as the rules imply, those just above the cutoff were much more likely to be promoted than those just below it.[16]

15. There is a way to adjust these differences to obtain an estimate of the effect of retention. We discuss this later in the book.

16. Jenny Nagaoka and Melissa Roderick, "Ending Social Promotion: The Effects of Retention" (Consortium of Chicago School Research, Chicago, 2004).

Nagaoka and Roderick found that third graders who scored just below the cutoff showed somewhat more improvement in performance the following year but that most of this difference had disappeared by the end of two years. For sixth graders, those scoring just above the cutoff had larger gains one year later, and this difference was similar two years later.

Is this the end of the discussion? Should we now be convinced that retention has only transitory benefits for third graders and longer-lasting negative effects for sixth graders? We will talk about more global reasons for not drawing this conclusion later in this section, but for now there are two issues that must be addressed. The first is specific to the study and helps to underline the importance of examining studies in detail rather than merely summarizing results. The second addresses a more general issue with the approach.

The first point is that Chicago had high-stakes testing in the third, sixth, and eighth grades. Students who were retained in grade therefore faced a high-stakes test one year after being retained. We might expect them to take the exam more seriously than someone taking a low-stakes exam and therefore to perform better. It is plausible that the one-year difference overstates the gain to retention. Conversely, sixth graders who were promoted took another high-stakes test two years later, while most of the sixth graders who were not promoted were taking a low-stakes test. This is likely to mean that the harm from retention two years later was exaggerated for sixth graders.

The second problem is that even though students above and below the cutoff were similar, they were not identical. There are good reasons both for believing students who did particularly badly on a test would show more improvement than other students and for believing that they would fall further behind. How can we get around this problem?

One way is to look at what happens when the cutoff is not important. Although they did not focus on it, Nagaoka and Roderick showed that in 2000, when the cutoff used in the study was not in effect, those below the study cutoff showed very slightly less improvement over one year than those slightly above the study cutoff. Unfortunately, data for two years later were not available.

The second approach is to look at differences with those slightly more above and slightly more below the cutoffs. To return to our earlier example, we could look at the difference between those scoring 63 and those scoring 64 on the test and between those scoring 65 and those scoring 66 on the test. If the difference between those scoring 64 and those scoring 65 on the test was similar to these other two differences, we would conclude that retention had no effect. If the improvement in score was greater going from 64 to 65 than from 63 to 64 or from 65 to 66, we would conclude that retention hurt the students, and if it was smaller, we would conclude that it helped them.

Brian Jacob and Lars Lefgren used essentially this approach to look at the Chicago data.[17] They examined both reading and mathematics scores because students had to

17. Brian A. Jacob and Lars Lefgren, "Remedial Education and Student Achievement: A Regression-Discontinuity Analysis," *Review of Economics and Statistics* 86 (February 2004): 226–44.

pass a threshold in each. Their conclusions are similar but not identical to those reached by Nagaoka and Roderick. In the third grade, they found that retention had a large positive one-year effect on reading that was mostly gone by the end of two years. In math, the positive one-year effect was diminished but not completely gone after two years. In the sixth grade they found a small negative one-year effect in reading that increased in the second year. For math, there was no clear negative effect in either year.

6.3. What Should We Conclude? The results of the two Chicago studies are reasonably consistent. For those close to the cutoff, retention in grade has some positive but probably transitory effects on third graders and some negative and possibly transitory effects on sixth graders. Although we might make a case that we should continue to study the effects on third graders, surely at least for older students, these new studies support the quotation at the beginning of this section that describes retention as "thoroughly discredited."

Sorry, but we should not be so fast to agree with that conclusion. There are a large number of questions we can ask that might stop us from drawing that conclusion even about retention of sixth graders. Both studies looked at the effect of retention on performance at the same age. What would happen if we looked at performance in the same grade? Did sixth graders who were retained do better or worse on the eighth-grade test than similar students who were promoted? We do not know. Because the mastery of math for the two groups was similar when the retained students were in seventh grade and the promoted students were in eighth grade, it is a good bet that the retained students did better on the eighth-grade test, but we cannot be sure until we check. And it is very uncertain whether they did better or worse on the eighth-grade reading test.

And, to make matters more complex, suppose that we do the study and we conclude that, in the long run, students retained in sixth grade do better than their promoted peers at each grade level but worse than their promoted peers at each age level. What policy should we favor? Our evaluation becomes very complex. There is good evidence that being old for their grade makes students more likely to drop out.[18] So some students will get less education because they were retained in grade, and, therefore, based on the assumption at the beginning of the paragraph, leave school with fewer skills. Other students will not reduce their education, and under these assumptions, will leave school with more skills. But to acquire these additional skills, they will have spent an extra year in school at great cost to both themselves and to the public. Even if retention does increase the skills of these students, it may be a very cost-ineffective approach.

There are at least three additional reasons that an advocate of retention might adhere to that position in the light of this research. The first is that, by their nature, the

18. Joshua D. Angrist and Alan B. Krueger, "The Effect of Age at School Entry on Educational Attainment: An Application of Instrumental Variables with Moments from Two Samples," *Journal of the American Statistical Association* 87 (June 1992): 328–36, and Susan E. Mayer and David Knutson, "Does the Timing of School Affect How Much Children Learn?" in Susan E. Mayer and Paul E. Peterson, eds., *Earning and Learning: How Schools Matter* (Washington, DC: Brookings Institution Press, 1999), 79–102.

Nagaoka-Roderick and Jacob-Lefgren studies tell us only about the effect of retention on students with scores near the cutoff chosen by the Chicago Public Schools (CPS). Perhaps retention is helpful to some students but the CPS set the promotion bar too high.

The second is that we have looked only at the effect on students who were retained, but there are important incentive effects from high-stakes testing.[19] If students work harder in order to pass the exam, perhaps the benefit from the extra effort outweighs any harm from a retention policy. We do not know that the incentive effects are important (or even positive), but if they are, they may make the effects positive for some students and negative for others. Even if policy analysts agreed on all the effects, they might reach different conclusions.

Finally, someone could conclude that the Chicago experience suggests that retention is not generally bad but rather that it is a bad policy as practiced by the CPS. They could maintain that a more sophisticated retention policy with special programs for retained students or one that relied on multiple and better indicators of academic mastery of the material would produce positive results.

7. Concluding Remarks

At this point you may be feeling frustration. You want to know whether retention is good or bad. Are the critics of "social promotion" correct, or are they just promoting a policy that sounds good, putting politics ahead of the sound judgment of most education professionals?

Throughout this book, I will try to play the role of objective arbiter. I regret to tell you that we will often end up where we are with the retention versus social promotion debate. There is some research that points in a particular direction, but there is a good deal that we do not know. For the most part, in this book I have resisted the temptation to discuss research that I consider bad and to elaborate on its obvious weaknesses. I certainly point out the weaknesses of many of the studies that I do discuss, but I view that differently. Many of the weaknesses are simply unavoidable, and it is often better to rely on a weak lamp than on none at all. Still, there is ample ground for reasonable people to disagree, both because the available research leaves many unanswered questions and because some policy positions depend on value judgments.

I am concerned that, having learned that it will not teach you how to solve poverty and discrimination, you will stop reading this book and drop the course in which it is assigned. But if you resist that temptation, I truly believe that by the end of this book you will be better at evaluating social policy options, and because you care enough about poverty and discrimination to be reading this book, that is important.

Therefore, in the last chapter, I will drop my cloak of academic distance and outline my conclusions, based on the material in this book, about what policies we should pursue. But that chapter is the least important chapter in the book.

For now, because the subject is only marginally related to poverty and discrimination policy, let me appease you by addressing what I think we should conclude about

19. We discuss this issue in more detail in the chapter on education reform.

retention in grade and social promotion. My strongest conclusion is that it is time to design a randomized study of retention and social promotion. Under the impetus of the No Child Left Behind Act and the standards movement, the United States is currently moving rapidly in the direction of widespread use of high-stakes testing and ending social promotion. Enormous resources are being directed at such testing. To me, it is unthinkable that we would not be devoting substantial resources to determining the effectiveness of the policy.

This may seem like a cheap conclusion for a researcher to reach, but it is not. There are strong moral restrictions on the types of experiments to which we should subject children. If I believed that retention was "thoroughly discredited," advocating experimental research would be immoral, as it would be if I believed the opposite. But I do believe that it is possible that we have been too lax historically (although the public probably underestimates the extent to which retention has become more common over the past twenty years).[20] I also believe that even if retention in grade proves less harmful than its critics maintain, it is unlikely to be a cost-effective policy for providing remediation. If we are moving in the direction of using retention more frequently, we should have better evidence of its effectiveness.

8. Further Reading

Anderson, Elijah. *Code of the Street.* New York: Norton, 1999.

Eaton, Susan E. *The Other Boston Busing Story.* New Haven, CT: Yale University Press, 2001.

Fuchs, Victor R., Alan B. Krueger, and James M. Poterba. "Economists' Views about Parameters, Values, and Policies: Survey Results in Labor and Public Economics." *Journal of Economic Literature* 36 (September 1998): 1387–425.

Kotlowitz, Alex. *There Are No Children Here.* New York: Doubleday, 1991.

Liebow, Elliot. *Tally's Corner.* Boston: Little, Brown and Company, 1967.

Lukas, J. Anthony. *Common Ground.* New York: Knopf, 1985.

9. Questions for Discussion

1. What is the difference between correlation and causality? If two events tend to occur together or sequentially, must one of them cause the other?
2. Explain what is meant by the Pareto principle. What are the limitations of this principle as a guide to policy?
3. A friend tells you that you should never be a patient in a teaching hospital because the death rate among patients in teaching hospitals is higher than in other hospitals. How do you respond?
4. Suppose that, relative to policy analysts who oppose more funding for job training programs, policy analysts who support more funding for job

20. See Robert M. Hauser, "Should We End Social Promotion? Truth and Consequences" (Working Paper 99-06, Center for Demography and Ecology, University of Wisconsin–Madison, 1999).

training programs, on average, believe that job training programs generate a larger increase in employment. Does this mean that policy analysts' beliefs are biased by their values?

5. Suppose that policy analysts who support and those who oppose more funding for Head Start, on average, have similar beliefs about the effectiveness of the program. Does this mean that policy analysts' views about policy are not influenced by their beliefs about the scientific evidence?

10. Appendix: A Quick Guide to Statistics

This appendix covers what you need to know to understand the statistics used in this book. It focuses on how to interpret the statistics rather than how they are calculated. It is not intended as a substitute for a standard statistics course that teaches the theory underlying the statistics.

10.1. Randomness. Before we discuss statistics, we need to think about what we mean when we say something is random. Suppose somebody shuffles a standard fifty-two-card deck. What is the probability that the top card is the ace of spades? In some sense, the probability is either one or zero. Either the top card is the ace of spades, in which case the probability is one, or the top card is not the ace of spades, in which case the probability is zero. But assuming that the person shuffled the deck fairly and that we have not looked at the top card or otherwise "cheated," from our perspective, the probability is 1/52. With more information, the outcome might not be random, but given our information it is.

The same will be true of many of the phenomena we study throughout this book. Test scores may or may not be random in some deep sense, but there are certainly many factors that affect test scores and that we do not measure. Therefore, from our perspective, test scores are random. Suppose we find one hundred pairs of students. We choose the pairs so that they look as similar as is feasible. To be part of the same pair, the students must be of the same sex, age, and race; have the same family structure; go to the same school; and have the same sixth-grade Iowa Test of Basic Skills math and reading scores. We will assume that under the rules of the school system, all are supposed to be retained in grade. As part of an experiment, one member of each pair is given a waiver from the school system rules and is promoted, while the other is retained in grade.

Unless promotion or retention is *much* better for all students (which seems unlikely), some of the "experimental" students who are promoted will do better and some will do worse than the "control" students who are retained in grade. Maybe one student in the pair was generally a good student but had been sick the night before the test that determined promotion or retention and did poorly. The other member of the pair was even weaker than the test scores suggest but had made a lot of lucky guesses. The first member of the pair will probably do better next time whether he is the one who is promoted or the one who is retained.

From our perspective, the number of promoted students who do better than their matched retained counterparts is a random variable. We could also look at the difference

between the test scores of the student who was promoted and the student who was retained. This, too, would be a random variable. And because each difference is a random variable, the average difference across the one hundred pairs is also a random variable.

10.2. The Mean and Standard Deviation. Suppose we give a test to a large number of students. We could list the entire distribution. That is, we could say that twenty got a 0, twenty-eight got a 1, thirty-five got a 2, and so on, up to the top score. If there are a lot of possible scores, this would be tedious and hard to interpret. One thing every student knows to ask is "What was the average on the test?" Statisticians call the average the *mean.*

But knowing the mean on the test is not enough. If the mean was 60, is a 50 an okay grade or a terrible grade? If the grades are all spread out from 0 to 100, a 50 is not too bad. If everyone got very close to 60, 50 is near the bottom. We would like to have a way of summarizing how dispersed the grades are without listing all the grades. There are many ways to do this.

One way statisticians measure dispersion is the standard deviation. The formula for the standard deviation is given in the next paragraph, but it is more important to understand how the standard deviation relates to the dispersion of the variable. We will discuss this relation after giving a formal definition of the standard deviation and describing the normal distribution.

To obtain the standard deviation, we first take all the observations of the random variable and calculate their mean. We then take the value of each observation of the random variable and subtract the mean. We then take this difference and multiply it by itself. This gives us the squared deviation of the measurement from the mean measurement. Next we add up all the squared deviations and divide by the number of measurements to get the mean squared deviation from the mean.[21] This is called the variance. The square root of the variance is the standard deviation.

10.3. The Normal Distribution. Many random variables have what is called a normal distribution. The normal distribution is sometimes referred to as a bell curve. The normal distribution has a very useful feature: it can be fully described by just two values, its mean and its standard deviation. We can use the mean and standard deviation to describe how likely it is that, if we pick randomly from a normal distribution, we will obtain a particular value of set of values. In particular, 95 percent of the time, a random variable drawn from a normal distribution will lie within 1.96 (or approximately two) standard deviations of the mean. Furthermore, 2.5 percent of the time it will be more than 1.96 standard deviations above the mean, and 2.5 percent of the time it will be more than 1.96 standard deviations below the mean. Similarly, 90 percent of the time it will be within 1.64 standard deviations of the mean, with the re-

21. We can also divide by the number of observations minus one. The measures have slightly different properties but for most practical purposes are indistinguishable.

maining 10 percent split equally between values more than 1.64 standard deviations above and below the mean.

10.4. Two Key Theorems. One important theorem in statistics says that (in most cases) if we take a lot of observations of a random variable and then take their mean, the estimated mean will be close to the true mean. So if we flip a coin a lot of times, the fraction of times that it comes up heads will be very close to the true probability of its coming up heads.

A second important theorem in statistics says that if we take the mean of a large number of independent random variables, the mean will be approximately normally distributed. Many random variables are the result of the offsetting effects of a large number of very small factors. For example, how tall someone is depends on both genetics and a large number of environmental factors. It is not surprising that we frequently observe the normal distribution in nature.

These two theorems are very helpful. They tell us, for example, that if we have enough pairs, the average difference in the test scores in our sample will be close to the true average difference and that the average will be approximately normally distributed.

Acting as if a distribution is normal often yields quite accurate results. It turns out that if we flip a fair coin one hundred times, the number of heads will have a standard deviation of five. Equivalently, the coin will come up heads an average of half (.5) of the time, with a standard deviation of .05. Using the normal distribution, we would expect the number of heads to be more than forty and less than sixty 95 percent of the time. We can show that the true probability is 94.3 percent, so the approximation is pretty accurate.

Of course, not all distributions are normal. Wages are not normally distributed, but if we take the logarithm of the wage, it is approximately normally distributed. Other distributions cannot be made normal even by redefining the variable. Despite this caveat, in most cases we will be quite accurate if we treat an estimate based on a large number of observations as normally distributed.

10.5. The Standard Error of an Estimate. Because an estimate is likely to be a normally distributed random variable, if we know its mean and standard deviation, we will know a great deal about it. In some cases, we can figure out the mean and standard deviation by relying on statistical theory. If promotion and retention are equally good, half the time the person who is promoted will do better and half the time the person retained will do better. If this is true, the mean number of our one hundred pairs in which the person promoted does better should be fifty. And, as mentioned earlier for the coin toss, the standard deviation will be five.

In other cases, we will have to estimate the standard deviation based on information about the observations in our data. Statisticians have developed methods for estimating what the standard deviation of the recorded mean will be without actually estimating lots of different means. We will not discuss exactly how they do this, but we can look at the general principle. If the test scores in our pairs differ by almost the same amount (e.g., in every pair the retained student does two points better than the

It is important to remember that whether or not they are reported along with the estimate, all estimates based on samples have associated standard errors. In this book and in newspapers, you will often find statements of this form: "According to the Bureau of the Census, the poverty rate in the United States in 2001 was 11.7 percent." In fact, the Census Bureau provides information to help readers calculate standard errors of its estimates.[22] Based on this information, we can calculate that the standard error of this estimate is about .14. We believe that 95 percent of the time if the Census Bureau had used a different (but similarly drawn) sample of the population, the estimated poverty rate would have fallen between 11.4 percent and 12.0 percent. For many purposes, it is unimportant whether the poverty rate is 11.7 percent or 11.4 percent, so being casual about standard errors causes no harm, but if we wish to make a point based on small differences in poverty rates, we will have to be aware that poverty rates are estimates and not exact. And we should remember that estimated poverty rates for smaller groups such as blacks or Hispanics are more imprecise.

promoted student), the standard deviation of the data will be low and our estimate of the standard deviation of the mean will also be low. If there is a lot of variation across pairs, so that in some pairs the retained student does a lot better and in others the promoted student does a lot better, the standard deviation of the data will be high, and so will the estimated standard deviation of the mean. We refer to our estimate of the standard deviation as the *standard error* of our estimate of the mean.

Thus, for example, we might report that our estimate of the difference in achievement between retained and promoted students was two points with a standard error of four. This is often written with the estimate on top and the standard error in parentheses underneath: $\begin{matrix} 2 \\ (4) \end{matrix}$.

10.6. Confidence Intervals and Statistical Significance. Our best estimate of the average of all the means we would record is the one mean we have actually calculated. Using the fact that the distribution of the mean is approximately normal, we estimate that if we were to estimate the mean many times, 95 percent of the time the estimated mean would lie within two standard errors of our estimate of the mean. That is a lot of "estimates," and it is important to keep this in mind when we look at real data.

In our earlier example, the students who were retained did better on average than those who were promoted. We would like to know how likely it is that we would obtain the same result if we did the experiment over again. Recall that our estimate of the average difference is that retained students do two points better than promoted stu-

22. U.S. Census Bureau, "Source and Accuracy of Estimates for Poverty in the United States: 2001," appendix to *Poverty in the United States: 2001,* Current Population Report P60-219 (Washington, DC: Government Printing Office, 2002).

dents. Our estimate of the standard error is four, and the distribution of our estimate is approximately normal. Therefore, we estimate that if we repeated the experiment, 95 percent of the time our estimate would lie between 2 – 1.96 standard errors or 2 – 1.96 × 4, or about –6, and 2 + 1.96 standard errors, or about 10. This range is called the 95 percent confidence interval because we believe (or are confident) that if we repeated the exercise, 95 percent of the time we would obtain an estimate in this range. To determine the 90 percent confidence interval, we would multiply the standard error by 1.64 instead of 1.96.

We could turn the question on its head by asking what our estimated standard error would be if the true average difference between the retained and promoted students were zero. Suppose that the estimated standard error in this case were also four. We would know that if the true difference were zero, 95 percent of the time we would obtain an estimate between –8 and 8. The probability of obtaining an estimate outside this range is 5 percent, or .05.

In practice, it usually makes little difference whether we ask whether two lies outside the confidence interval we would have if the true value were zero or whether zero lies outside the 95 percent confidence interval based on our estimate of a difference of two.[23] Because it is usually simpler, we more frequently ask whether zero is outside the 95 percent confidence interval based on our estimate of two, but we conclude that if the true value were zero, it is (un)likely that we would obtain an estimate of two.

10.7. Statistical Significance. In statistical jargon, we say that the difference between our estimate and some value is statistically significant if the value lies outside the 95 percent confidence interval. Because the probability of something outside the 95 percent confidence interval is 5 percent, we will say that the difference is statistically significant at the 5 percent or .05 level. We may also decide to use a different confidence interval, 90 percent or 99 percent, in which case we will say that the difference between our estimate and the value is significant at the .1 or the .01 level. If some value falls within the confidence interval of our estimate, we will say that the difference between our estimate and that value is statistically insignificant. Whether our estimate is statistically significantly or insignificantly different from a value depends on the level of significance that we choose. The difference may be statistically significant at the .1 level but insignificant at the .05 level (but, of course, not the reverse). In the earlier example, our estimate of a two-point difference is statistically insignificant whether we choose the .05 or the .1 level.

It is important to recognize that statistical significance does not mean statistical importance. It is unfortunate that statisticians adopted the word "significant." Suppose we had a very large sample of wages of men and women and we estimated that, relative to men, women on average earned five dollars less per year. If the standard error of this

23. The rationale for this is as follows. If the true difference is zero, our estimated difference will be close to zero and the estimated standard error will be very similar. If the estimate is very different from zero, the estimated standard error may be quite different, but because the estimate is a long way from zero, in either case, the estimate and zero will fall outside each other's confidence interval.

estimate were only two dollars, a difference of zero dollars would lie outside the 95 percent confidence interval of one to nine, and the difference would be statistically significant but probably of no social significance whatsoever. On the other hand, with a small sample, a large estimated difference may not be statistically significant. In this case, if correct, the difference might be socially important, but because of the small sample, we have little confidence in the precision of our estimate. The true difference might be much larger, or there might be no difference whatsoever.

Statistics texts used to say that before performing a test, a researcher should choose the significance level and then report whether the effect being studied was statistically significant at that level. This approach is problematic. First, if one researcher chose a significance level of .1 and another a significance level of .05, they could conduct identical experiments and draw different conclusions. In addition, the reader had no way of verifying that the researcher chose the significance level before learning the results and often suspected that the significance level was chosen ex post facto on the basis that best suited the researcher.

Moreover, recall that we do not usually know the standard error and must estimate it. Therefore, we only have an estimate of how improbable an estimate is. Perhaps most important, our goal is to ask a question like "Do poverty rates differ between blacks and whites?" Our answer takes the form of a probability based on the data: "If there were no difference in the poverty rates of blacks and whites, the probability of finding a difference this large would be less than [a given number]." Our conclusion about whether blacks and whites have different poverty rates should not differ substantially if that number is 5.0001 percent or 4.9999 percent. Therefore, it is best to think about the significance level of a difference and not focus too much on whether it is above or below some critical value.

10.8. The *t*-Statistic. If we ask whether some value lies within the x percent confidence interval for our estimate, we are asking whether

$$\text{estimate} + t^* \times \text{standard error} > \text{value}$$

and

$$\text{value} > \text{estimate} - t^* \times \text{standard error},$$

where t^* is the value that determines the size of the confidence interval. If we were interested in the 95 percent confidence interval, t^* would be 1.96. For the 90 percent confidence interval, it would be 1.64.

A little algebra shows that asking whether the value is in the confidence interval is the same as asking whether

$$\text{absolute value}\left(\frac{\text{estimate} - \text{value}}{\text{standard error}}\right) < t^*.$$

Thus, if we want to know whether our two-point difference lies within the 95 percent confidence interval, we divide it by the estimated standard error (four):

$$\text{absolute value}\left(\frac{\text{estimate} - \text{value}}{\text{standard error}}\right) = \text{absolute value}\left(\frac{2-0}{4}\right) = .5.$$

Because .5 is less than 1.96, two lies inside the 95 percent confidence interval, and our estimate of two is not significantly different from zero at the .05 level. Indeed, it is not significant at the .1 level.

Note that in order to determine whether the difference is significant, all we have to do is divide the difference by the standard error. This ratio is called the *t*-statistic. We then compare the absolute value of the *t*-statistic to our chosen critical value, or we can report the significance level of the *t*-statistic based on statistical tables. For most purposes, it is sufficient to remember that the probability of a *t*-statistic greater in absolute value than 1.64 is about .1 and the probability of a *t*-statistic greater in absolute value than 1.96 is about .05 and that the probability declines rapidly as the absolute value of the *t*-statistic exceeds two.

Often, as in our example, we are interested in whether some estimate equals zero. Is the difference in earnings between men and women statistically significantly different from zero? In this case, our *t*-statistic becomes

$$t\text{-statistic} = \left(\frac{\text{estimate}}{\text{standard error}}\right).$$

We have noted that our tables will frequently present an estimate with its standard error in parentheses underneath. This allows us to calculate the *t*-statistic quickly. Some authors use the same format but report the *t*-statistic instead of the standard error in parentheses. Be careful to check which convention is used when reading different sources.

10.9. Relations among Variables. Often we are interested in questions that relate one variable to another. We implicitly ask such questions when we ask about differences among groups, such as "Is the poverty rate higher for blacks than for whites?" But we may be interested in questions like "How does the poverty rate vary with the state of the economy?"

We might observe the poverty rate and the unemployment rate (a measure of the state of the economy) over a period of years. We can plot these combinations of poverty and unemployment rates on a two-dimensional diagram as in figure 1.1. This gives us a good visual sense of whether there is a relation between unemployment and the poverty rate, but we still require some way of summarizing the relation. We would like to say something like "For each percentage point increase in the unemployment rate, the poverty rate increases by *x* percentage points." To summarize the data, we can fit a line to the points. Obviously a straight line will not fit all of the points perfectly, but we can choose the line on the basis of how well it fits. Figure 1.1 fits one possible line.

Statistics courses focus on different techniques for choosing the best line. For this book, you will not need to know how to fit a line or the various advantages and disadvantages of different techniques. You will need to know how to read a table showing

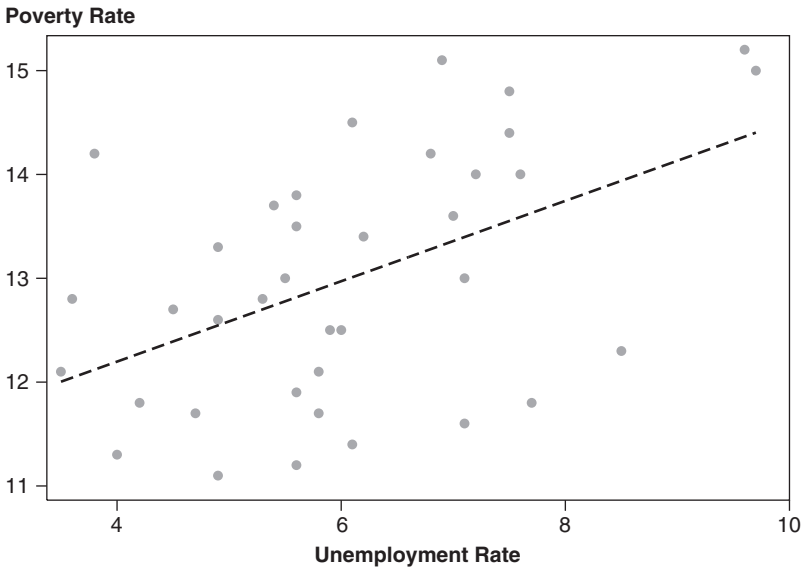

Figure 1.1 Relation between Poverty Rate and Unemployment Rate

the results of the line-fitting exercise. You will also need to remember that it is the analyst, not the data, who argues the direction of causality.

The results of the line-fitting in figure 1.1 can be summarized in a table that describes the line we have fitted. Table 1.1 is presented in a way similar to the estimates we presented above. The first column in the table (the one headed "Without Median Male Earnings") says that the poverty rate increases by about .39 percentage points for every one percentage point increase in the unemployment rate.

The coefficient, .39, is only an estimate of the relation between the unemployment rate and the poverty rate. It is based on a sample of years. If we had chosen different years, we might have obtained different answers.

Recall that we will be much more confident of the precision of our estimate of the difference in test scores if they are clustered around the same value than if they vary substantially. Similarly, we will be more confident of the precision of our coefficient estimate if the points are clustered around our fitted line.

As with the mean, we summarize the precision of our coefficient estimate by the standard error. If the coefficient estimate is normally distributed (and theory shows that it generally will be), if we looked at a different 1967 through 2000, 95 percent of the time we would obtain an estimate within 1.96 standard errors of .39.

In our example, the standard error is about .12. The 95 percent confidence interval therefore runs from about $.39 - 1.96 \times .12$, or about .15, to $.39 + 1.96 \times .12$, or about .63. We can divide the coefficient by the standard error to obtain a *t*-statistic of about three. It is very unlikely that if the true coefficient were zero we would have obtained a *t*-statistic this large in absolute value by chance. We are therefore reasonably confident that the relation is not due to random sampling error.

Table 1.1 Relation between Poverty Rate and Unemployment Rate (Sample Table)

	Without Median Male Earnings	*With Median Male Earnings*
Unemployment Rate	0.39	0.15
	(0.12)	(0.12)
Median Male Earnings (thousands of dollars)	—	–0.47
		(0.12)
Constant	10.65	26.17
	(0.73)	(4.10)

Note: Standard errors are in parentheses.

It is important to remember that finding that a relation is unlikely to be the result of random sampling does not establish that it is causal. As discussed in the introduction to this book, we have merely established that the poverty rate and the unemployment rate have tended to change in the same direction. The poverty rate could be changing the unemployment rate, or both rates could be influenced by some other factor.

10.10. Controlling for Other Factors. In our fictional example earlier in this appendix, we assumed that we were able to match students on the basis of their sex, age, race, family structure, school, and sixth-grade ITBS math and reading scores. In practice, we are unable to match people exactly. We need techniques that allow us to ask an "if" question such as "What would the effect of promotion be *if* two individuals were identical in all these dimensions?"

If we think that some other factor might account for the relation between the unemployment rate and the poverty rate, we want to ask, "What would be the effect on the poverty rate of an increase in the unemployment rate if this other factor did not change?" For example, there is some evidence that wages fluctuate over the business cycle, so periods of high unemployment might also be periods of low wages.[24] Are the changes in the poverty rate driven by changes in the unemployment rate, prevailing wages, or both?

Conceptually, to answer this question we would like to compare periods with differing unemployment rates in which prevailing wages were constant and also to compare the poverty rates in periods with differing prevailing wages but similar unemployment rates. Of course, such perfect correspondence may not occur in the data. However, statistical techniques allow us to do something comparable.

The right-hand column of table 1.1 shows the results of one such technique. Each coefficient should be interpreted as the effect on the poverty rate of varying that factor

24. Gary Solon, Robert Barsky, and Jonathan A. Parker, "Measuring the Cyclicality of Real Wages: How Important Is Composition Bias?" *Quarterly Journal of Economics* 109 (February 1994): 1–25.

while holding the other factors constant. Thus, the results show that, holding median male earnings constant, a one-point increase in the unemployment rate raises the poverty rate by about .15 percentage points. The standard error of this estimate is .12, which gives a *t*-statistic of

$$\frac{.15}{.12} = 1.25,$$

well below 1.96 or even 1.64. Obtaining a *t*-statistic of this magnitude is quite likely even if the true effect of the unemployment rate on the poverty rate is zero. We therefore do not have any good evidence that the unemployment rate affects the poverty rate.

On the other hand, even holding the unemployment rate constant, increases in median male earnings are associated with quite noticeable reductions in the poverty rate. If we were to compare two periods with the same unemployment rate in one of which median male earnings exceeded those in the other period by $1,000, we would expect that, on average, the period with the higher earnings would have a poverty rate about .5 percentage points lower than the period with the lower earnings.

The standard error of this estimate is also .12, so the *t*-statistic is close to four. It is very unlikely that we would observe a *t*-statistic of this magnitude if there were no real relation between median male earnings and the poverty rate.

Of course, the relation between median male earnings and the poverty rate, conditional on the unemployment rate, may still be due to some other factor. We address this issue in greater depth elsewhere in this book.