1 Introduction: why natural experiments?

> If I had any desire to lead a life of indolent ease, I would wish to be an identical twin, separated at birth from my brother and raised in a different social class. We could hire ourselves out to a host of social scientists and practically name our fee. For we would be exceedingly rare representatives of the only really adequate natural experiment for separating genetic from environmental effects in humans—genetically identical individuals raised in disparate environments.
>
> —Stephen Jay Gould (1996: 264)

Natural experiments are suddenly everywhere. Over the last decade, the number of published social-scientific studies that claim to use this methodology has more than tripled (Dunning 2008a). More than 100 articles published in major political-science and economics journals from 2000 to 2009 contained the phrase "natural experiment" in the title or abstract—compared to only 8 in the three decades from 1960 to 1989 and 37 between 1990 and 1999 (Figure 1.1).[1] Searches for "natural experiment" using Internet search engines now routinely turn up several million hits.[2] As the examples surveyed in this book will suggest, an impressive volume of unpublished, forthcoming, and recently published studies—many not yet picked up by standard electronic sources—also underscores the growing prevalence of natural experiments.

This style of research has also spread across various social science disciplines. Anthropologists, geographers, and historians have used natural experiments to study topics ranging from the effects of the African slave trade to the long-run consequences of colonialism. Political scientists have explored the causes and consequences of suffrage expansion, the political effects of military conscription, and the returns to campaign donations. Economists, the most prolific users of natural experiments to date, have scrutinized the workings of

[1] Such searches do not pick up the most recent articles, due to the moving wall used by the online archive, JSTOR.

[2] See, for instance, Google Scholar: http://scholar.google.com.

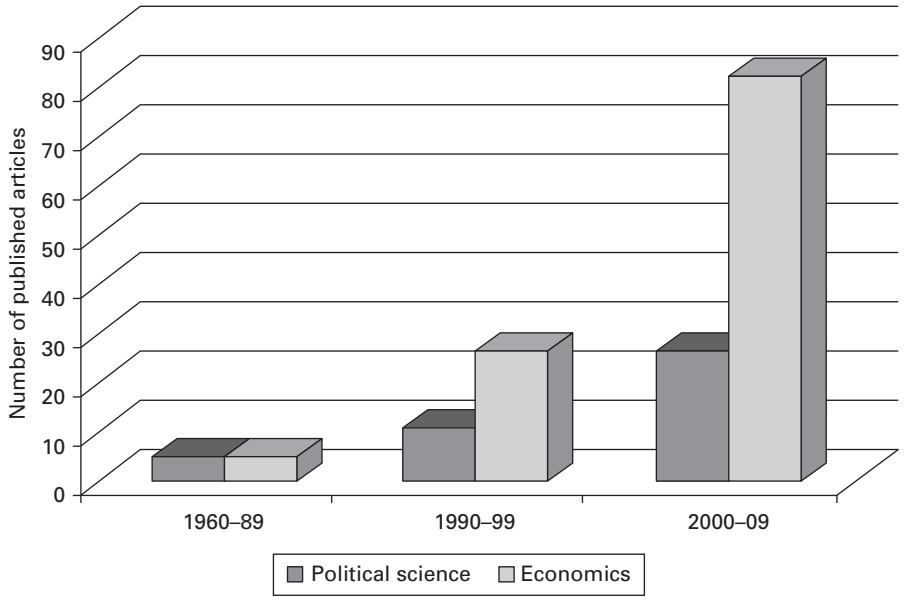

Figure 1.1 Natural experiments in political science and economics
Articles published in major political science and economics journals with "natural experiment" in the title or abstract (as tracked in the online archive JSTOR).

labor markets, the consequences of schooling reforms, and the impact of institutions on economic development.[3]

The ubiquity of this method reflects its potential to improve the quality of causal inferences in the social sciences. Researchers often ask questions about cause and effect. Yet, those questions are challenging to answer in the observational world—the one that scholars find occurring around them. Confounding variables associated both with possible causes and with possible effects pose major obstacles. Randomized controlled experiments offer one possible solution, because randomization limits confounding. However, many causes of interest to social scientists are difficult to manipulate experimentally.

Thus stems the potential importance of natural experiments—in which social and political processes, or clever research-design innovations, create

[3] According to Rozenzweig and Wolpin (2000: 828), "72 studies using the phrase 'natural experiment' in the title or abstract issued or published since 1968 are listed in the *Journal of Economic Literature* cumulative index." A more recent edited volume by Diamond and Robinson (2010) includes contributions from anthropology, economics, geography, history, and political science, though several of the comparative case studies in the volume do not meet the definition of natural experiments advanced in this book. See also Angrist and Krueger (2001), Dunning (2008a, 2010a), Robinson, McNulty, and Krasno (2009), Sekhon (2009), and Sekhon and Titiunik (2012) for surveys and discussion of recent work.

situations that approximate true experiments. Here, we find observational settings in which causes are randomly, or as good as randomly, assigned among some set of units, such as individuals, towns, districts, or even countries. Simple comparisons across units exposed to the presence or absence of a cause can then provide credible evidence for causal effects, because random or as-if random assignment obviates confounding. Natural experiments can help overcome the substantial obstacles to drawing causal inferences from observational data, which is one reason why researchers from such varied disciplines increasingly use them to explore causal relationships.

Yet, the growth of natural experiments in the social sciences has not been without controversy. Natural experiments can have important limitations, and their use entails specific analytic challenges. Because they are not so much planned as discovered, using natural experiments to advance a particular research agenda involves an element of luck, as well as an awareness of how they have been used successfully in disparate settings. For natural experiments that lack true randomization, validating the definitional claim of as-if random assignment is very far from straightforward. Indeed, the status of particular studies as "natural experiments" is sometimes in doubt: the very popularity of this form of research may provoke conceptual stretching, in which an attractive label is applied to research designs that only implausibly meet the definitional features of the method (Dunning 2008a). Social scientists have also debated the analytic techniques appropriate to this method: for instance, what role should multivariate regression analysis play in analyzing the data from natural experiments? Finally, the causes that Nature deigns to assign at random may not always be the most important causal variables for social scientists. For some observers, the proliferation of natural experiments therefore implies the narrowing of research agendas to focus on substantively uninteresting or theoretically irrelevant topics (Deaton 2009; Heckman and Urzúa 2010). Despite the enthusiasm evidenced by their increasing use, the ability of natural experiments to contribute to the accumulation of substantively important knowledge therefore remains in some doubt.

These observations raise a series of questions. How can natural experiments best be discovered and leveraged to improve causal inferences in the service of diverse substantive research agendas? What are appropriate methods for analyzing natural experiments, and how can quantitative and qualitative tools be combined to construct such research designs and bolster their inferential power? How should we evaluate the success of distinct natural experiments, and what sorts of criteria should we use to assess their strengths and limitations? Finally, how can researchers best use natural experiments to

build strong research designs, while avoiding or mitigating the potential limitations of the method? These are the central questions with which this book is concerned.

In seeking to answer such questions, I place central emphasis on natural experiments as a "design-based" method of research—one in which control over confounding variables comes primarily from research-design choices, rather than *ex post* adjustment using parametric statistical models. Much social science relies on multivariate regression and its analogues. Yet, this approach has well-known drawbacks. For instance, it is not straightforward to create an analogy to true experiments through the inclusion of statistical controls in analyses of observational data. Moreover, the validity of multivariate regression models or various kinds of matching techniques depends on the veracity of causal and statistical assumptions that are often difficult to explicate and defend—let alone validate.[4] By contrast, random or as-if random assignment usually obviates the need to control statistically for potential confounders. With natural experiments, it is the research design, rather than the statistical modeling, that compels conviction.

This implies that the quantitative analysis of natural experiments can be simple and transparent. For instance, a comparison of average outcomes across units exposed to the presence or absence of a cause often suffices to estimate a causal effect. (This is true at least in principle, if not always in practice; one major theme of the book is how the simplicity and transparency of statistical analyses of natural experiments can be bolstered.) Such comparisons in turn often rest on credible assumptions: to motivate difference-of-means tests, analysts need only invoke simple causal and statistical models that are often persuasive as descriptions of underlying data-generating processes.

Qualitative methods also play a critical role in natural experiments. For instance, various qualitative techniques are crucial for discovering opportunities for this kind of research design, for substantiating the claim that assignment to treatment variables is really as good as random, for interpreting, explaining, and contextualizing effects, and for validating the models used in quantitative analysis. Detailed qualitative information on the circumstances that created a natural experiment, and especially on the process by which "nature" exposed or failed to expose units to a putative cause, is often essential. Thus, substantive and contextual knowledge plays an important role at every

[4] Matching designs, including exact and propensity-score matching, are discussed below. Like multiple regression, such techniques assume "selection on observables"—in particular, that unobserved confounders have been measured and controlled.

stage of natural-experimental research—from discovery to analysis to evaluation. Natural experiments thus typically require a mix of quantitative and qualitative research methods to be fully compelling.

In the rest of this introductory chapter, I explore these themes and propose initial answers to the questions posed above, which the rest of the book explores in greater detail. The first crucial task, however, is to define this method and distinguish it from other types of research designs. I do this below, after first discussing the problem of confounding in more detail and introducing several examples of natural experiments.

1.1 The problem of confounders

Consider the obstacles to investigating the following hypothesis, proposed by the Peruvian development economist Hernando de Soto (2000): granting *de jure* property titles to poor land squatters augments their access to credit markets, by allowing them to use their property to collateralize debt, thereby fostering broad socioeconomic development. To test this hypothesis, researchers might compare poor squatters who possess titles to those who do not. However, differences in access to credit markets across these groups could in part be due to confounding factors—such as family background—that also make certain poor squatters more likely to acquire titles to their property.

Investigators may seek to control for such confounders by making comparisons between squatters who share similar values of confounding variables but differ in their access to land titles. For instance, a researcher might compare titled and untitled squatters with parallel family backgrounds. Yet, important difficulties remain. First, the equivalence of family backgrounds is difficult to assess: for example, what metric of similarity should be used? Next, even supposing that we define an appropriate measure and compare squatters with equivalent family backgrounds, there may be other difficult-to-measure confounders—such as determination—that are associated with obtaining titles and that also influence economic and political behaviors. Differences between squatters with and without land titles might then be due to the effect of the titles, the effect of differences in determination, or both.

Finally, even if confounders *could* all be identified and successfully measured, the best way to "control" for them is not obvious. One possibility is stratification, as mentioned above: a researcher might compare squatters who have equivalent family backgrounds and measured levels of determination—but who vary with respect to whether or not they have land titles. However,

such stratification is often infeasible, among other reasons because the number of potential confounders is usually large relative to the number of data points (that is, relative to the number of units).[5] A cross-tabulation of titling status against every possible combination of family background and levels of determination would be likely to have many empty cells. For instance, there may be no two squatters with precisely the same combination of family attributes, such as parental education and income, and the same initial determination, but different exposures to land titles.

Analysts thus often turn to conventional quantitative methods, such as multivariate regression or its analogues, to control for observable confounders. The models essentially extrapolate across the missing cells of the cross-tabulations, which is one reason for their use. Yet, typical regression models rely on essentially unverifiable assumptions that are often difficult to defend. As I discuss in this book, this is an important difficulty that goes well beyond the challenge of identifying and measuring possible confounders.

1.1.1 The role of randomization

How, then, can social scientists best make inferences about causal effects? One option is true experimentation. In a randomized controlled experiment to estimate the effects of land titles, for instance, some poor squatters might be randomly assigned to receive *de jure* land titles, while others would retain only de facto claims to their plots. Because of randomization, possible confounders such as family background or determination would be balanced across these two groups, up to random error (Fisher [1935] 1951). After all, the flip of a coin determines which squatters get land titles. Thus, more determined squatters are just as likely to end up without titles as with them. This is true of all other potential confounders as well, including family background. In sum, randomization creates *statistical independence* between these confounders and treatment assignment—an important concept discussed later in the book.[6] Statistical independence implies that squatters who are likely to do poorly even if they are granted titles are initially as likely to receive them as not to receive them. Thus, particularly when the number of squatters in each group is large and so the role of random error is small, squatters with titles and without titles should be nearly indistinguishable as groups—save for the

[5] This stratification strategy is sometimes known as "exact matching." One reason exact matching may be infeasible is that covariates—that is, potential confounders—are continuous rather than discrete.

[6] In Chapter 5, when I introduce the idea of *potential outcomes*, I discuss how randomization creates statistical independence of potential outcomes and treatment assignment.

presence or absence of titles. *Ex post* differences in outcomes between squatters with and without land titles are then most likely due to the effect of titling.

In more detail, random assignment ensures that any differences in outcomes between the groups are due either to chance error or to the causal effect of property titles. In any one experiment, of course, one or the other group might end up with more determined squatters, due to the influence of random variation; distinguishing true effects from chance variation is the point of statistical hypothesis testing (Chapter 6). Yet, if the experiment were to be repeated over and over, the groups would not differ, on average, in the values of potential confounders. Thus, the average of the average difference of group outcomes, across these many experiments, would equal the true difference in outcomes—that is, the difference between what would happen if every squatter were given titles, and what would happen if every squatter were left untitled. A formal definition of this causal effect, and of estimators for the effect, will await Chapter 5. For now, the key point is that randomization is powerful because it obviates confounding, by creating *ex ante* symmetry between the groups created by the randomization. This symmetry implies that large post-titling differences between titled and untitled squatters provide reliable evidence for the causal effect of titles.

True experiments may offer other advantages as well, such as potential simplicity and transparency in the data analysis. A straightforward comparison, such as the difference in average outcomes in the two groups, often suffices to estimate a causal effect. Experiments can thus provide an attractive way to address confounding, while also limiting reliance on the assumptions of conventional quantitative methods such as multivariate regression—which suggests why social scientists increasingly utilize randomized controlled experiments to investigate a variety of research questions (Druckman et al. 2011; Gerber and Green 2012; Morton and Williams 2010).

Yet, in some contexts direct experimental manipulation is expensive, unethical, or impractical. After all, many of the causes in which social scientists are most interested—such as political or economic institutions—are often not amenable to manipulation by researchers. Nor is true randomization the means by which political or economic institutions typically allocate scarce resources. While it is not inconceivable that policy-makers might roll out property titles in a randomized fashion—for example, by using a lottery to determine the timing of titling—the extension of titles and other valued goods typically remains under the control of political actors and policy-makers (and properly so). And while examples of randomized interventions are becoming more frequent (Gerber and Green 2012), many other causes continue to be

allocated by social and political process, not by experimental researchers. For scholars concerned with the effects of causes that are difficult to manipulate, natural experiments may therefore provide a valuable alternative tool.

1.2 Natural experiments on military conscription and land titles

In some natural experiments, policy-makers or other actors do use lotteries or other forms of true randomization to allocate resources or policies. Thus, while the key intervention is not planned and implemented by an experimental researcher—and therefore these are observational studies, not experiments—such randomized natural experiments share with true experiments the attribute of randomized assignment of units to "treatment" and "control" groups.[7]

For instance, Angrist (1990a) uses a randomized natural experiment to study the effects of military conscription and service on later labor-market earnings. This topic has important social-scientific as well as policy implications; it was a major source of debate in the United States in the wake of the Vietnam War. However, the question is difficult to answer with data from standard observational studies. Conscripted soldiers may be unlike civilians; and those who volunteer for the military may in general be quite different from those who do not. For example, perhaps soldiers volunteer for the army because their labor-market prospects are poor to begin with. A finding that ex-soldiers earn less than nonsoldiers is then hardly credible evidence for the effect of military service on later earnings. Confounding factors—those associated with both military service and economic outcomes—may be responsible for any such observed differences.

From 1970 to 1972, however, the United States used a randomized lottery to draft soldiers for the Vietnam War. Cohorts of 19- and 20-year-old men were randomly assigned lottery numbers that ranged from 1 to 366, according to their dates of birth. All men with lottery numbers below the highest number called for induction each year were "draft eligible," while those with higher numbers were not eligible for the draft. Using earnings records from the Social Security Administration, Angrist (1990a) estimates modest negative effects of draft eligibility on later income. For example, among white men who were

[7] I use the terms "independent variable," "treatment," and "intervention" roughly synonymously in this book, despite important differences in shades of meaning. For instance, "intervention" invokes the idea of manipulability—which plays a key role in many discussions of causal inference (e.g., Holland 1986)—much more directly than "independent variable."

eligible for the draft in 1971, average earnings in 1984 were $15,813.93 in current US dollars, while in the ineligible group they were $16,172.25. Thus, assignment to draft eligibility in 1971 caused an estimated decrease in average yearly earnings of $358.32, or about a 2.2 percent drop from average earnings of the assigned-to-control group.[8]

The randomized natural experiment plays a key role in making any causal inferences about the effects of military conscription persuasive. Otherwise, initial differences in people who were or were not drafted could explain any *ex post* differences in economic outcomes or political attitudes.[9] The usefulness of the natural experiment is that confounding should not be an issue: the randomization of draft lottery ensures that on average, men who were draft eligible are just like those who were not. Thus, large *ex post* differences are very likely due to the effects of the draft.

Of course, in this case not all soldiers who were drafted actually served in the military: some were disqualified by physical and mental exams, some went to college (which typically deferred induction during the Vietnam War), and others went to Canada. By the same token, some men who were not drafted volunteered. It might therefore seem natural to compare the men who actually served in the military to those who did not. Yet, this comparison is again subject to confounding: soldiers self-select into military service, and those who volunteer are likely different in ways that matter for earnings from those who do not. The correct, natural-experimental comparison is between men randomly assigned to draft eligibility—whether or not they actually served—and the whole assigned-to-control group. This is called "intention-to-treat" analysis—an important concept I discuss later in this book.[10] Intention-to-treat analysis estimates the effect of draft eligibility, not the effect of actual military service. Under certain conditions, the natural experiment can also be used to estimate the effects of draft eligibility on men who would serve if drafted, but otherwise would not.[11] This is the goal of instrumental-variables analysis, which is discussed later in this book—along with the key assumptions that must be met for its persuasive use.

Not all natural experiments feature a true randomized lottery, as in Angrist's study. Under some conditions, social and political processes may

[8] The estimate is statistically significant at standard levels; see Chapters 4 and 6.
[9] An interesting recent article by Erikson and Stoker (2011) uses this same approach to estimate the effects of draft eligibility on political attitudes and partisan identification.
[10] See Chapters 4 and 5.
[11] These individuals are called "Compliers" because they comply with the treatment condition to which they are assigned (Chapter 5).

assign units to treatment and control groups in a way that is persuasively *as-if* random. In such settings, ensuring that confounding variables do not distort results is a major challenge, since no true randomizing device assigns units to the treatment and control groups. This is one of the main challenges—and sometimes one of the central limitations—of much natural-experimental research, relative for instance to true experiments. Yet, social or political processes, or clever research-design innovations, sometimes do create such opportunities for obviating confounding. How to validate the claim that assignment to comparison groups is plausibly as good as random in such studies is an important focus of this book.

Galiani and Schargrodsky (2004, 2010) provide an interesting example on the effects of extending property titles to poor squatters in Argentina. In 1981, squatters organized by the Catholic Church occupied an urban wasteland in the province of Buenos Aires, dividing the land into similar-sized parcels that were then allocated to individual families. A 1984 law, adopted after the return to democracy in 1983, expropriated this land, with the intention of transferring title to the squatters. However, some of the original owners then challenged the expropriation in court, leading to long delays in the transfer of titles to the plots owned by those owners, while other titles were ceded and transferred to squatters immediately.

The legal action therefore created a "treatment" group—squatters to whom titles were ceded immediately—and a "control" group—squatters to whom titles were not ceded.[12] Galiani and Schargrodsky (2004, 2010) find significant differences across these groups in subsequent housing investment, household structure, and educational attainment of children—though not in access to credit markets, which contradicts De Soto's theory that the poor will use titled property to collateralize debt. They also find a positive effect of property rights on self-perceptions of individual efficacy. For instance, squatters who were granted land titles—for reasons over which they apparently had no control!—disproportionately agreed with statements that people get ahead in life due to hard work (Di Tella, Galiani, and Schargrodsky 2007).

Yet, what makes this a natural experiment, rather than a conventional observational study in which squatters with and without land titles are compared? The key definitional criterion of a natural experiment, as we shall see below, is that the assignment of squatters to treatment and control

[12] I use the terms "treatment" and "control" groups here for convenience, and by way of analogy to true experiments. There is no need to define the control group as the absence of treatment, though in this context the usage makes sense (as we are discussing the presence and absence of land titles). One could instead talk about "treatment group 1" and "treatment group 2," for example.